Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet, pages 8–12
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Finding Corresponding Constructions in English and Japanese in a TED Talk Parallel Corpus using Frames-and-Constructions Analysis

Kyoko Ohara
Keio University/RIKEN
4-1-1 Hiyoshi, Kohoku-ku, Yokohama City 223-8521, Japan
ohara@hc.st.keio.ac.jp

Abstract

This paper reports on an effort to search for corresponding constructions in English and Japanese in a TED Talk parallel corpus, using frames-and-constructions analysis (Ohara, 2019; Ohara and Okubo, 2020; cf. Czulo, 2013, 2017). The purpose of the paper is two-fold: (1) to demonstrate the validity of frames-and-constructions analysis to search for corresponding constructions in typologically unrelated languages; and (2) to assess whether the "Do schools kill creativity?" TED Talk parallel corpus, annotated in various languages for Multilingual FrameNet, is a good starting place for building a multilingual constructicon. The analysis showed that similar to our previous findings involving texts in a Japanese to English bilingual children's book, the TED Talk bilingual transcripts include pairs of constructions that share similar pragmatic functions. While the TED Talk parallel corpus constitutes a good resource for frame semantic annotation in multiple languages, it may not be the ideal place to start aligning constructions among typologically unrelated languages. Finally, this work shows that the proposed method, which focuses on heads of sentences, seems valid for searching for corresponding constructions in transcripts of spoken data, as well as in written data of typologically-unrelated languages.

Keywords: Japanese FrameNet, pragmatic function, multilingual constsructicon

1. Introduction

This paper reports on an effort to find corresponding Japanese and English grammatical constructions in a TED Talk parallel corpus, using the frames-and-constructions analysis method proposed in Ohara (2019) and Ohara and Okubo (2020). The method focuses on heads of sentences in language, where a head is defined as "the most contentful word that most closely denotes the same function as the phrase (or clause) as a whole (cf. Croft, In Preparation: 417)." The purpose of the paper is two-fold: (1) to demonstrate the validity of frames-and-construction analysis as a methodology to search for corresponding constructions in a pair of typologically-unrelated languages such as English and Japanese; and (2) to assess whether the "Do schools kill creativity?" TED Talk parallel corpus, whose sentences have been annotated in frame-semantic terms in various languages, including English, Brazilian Portuguese, French, German, and Japanese for Multilingual FrameNet, is a good starting place to align constructions for building a multilingual/contrastive constructicon.

Our analysis revealed the following:

- There are indeed pairs of sentences that constitute instances of corresponding constructions in English and Japanese that share similar pragmatic functions in the TED Talk bilingual transcripts, similar to our previous findings for texts in a Japanese–English bilingual children's book;
- While the TED Talk parallel corpus constitutes a good resource for frame semantic annotation, it may not be the ideal place to start aligning constructions across typologically-unrelated languages, likely as a result of characteristics of the genre of subtitles;
- The proposed frames-and-constructions analysis method, an approach that focuses on heads of sentences, seems valid to search for corresponding constructions in transcripts of spoken data, as well as in written data of typologically-unrelated languages.

The organization of the rest of the paper is as follows. Section 2 discusses background to the study. Section 3 presents the hypothesis, method, and the results of the analysis. Section 4 addresses the functional mismatches in the parallel corpus, the validity of the method, and the appropriateness of the corpus as a starting point for aligning constructions in a multilingual constructicon. Finally, Section 5 provides a conclusion and prospects for future work.

2. Related Work

The frames-and-constructions analysis method describes meanings and structures of sentences, focusing on the semantic frames evoked by various linguistic expressions in the sentences. It is grounded in the theories of Frame Semantics and Construction Grammar (Fillmore and Baker, 2010; Fillmore, 2013). Czulo (2013, 2017, elsewhere) proposed this method as a translation model, based on analyses of German and English parallel data[1]. Those works hypothesized that ideally the semantic frame of the translation matches that of the original (the primacy of the frame hypothesis). However, often cases of frame mismatches exist between pairs of source and target sentences and Czulo (2013, 2017) argued that structural divergence can be a cause for frame mismatch, in addition to cultural, typological, and perspectival differences. Czulo also observed that even when a frame mismatch exists because of structural divergence between source and target sentences, the two sentences may share the same pragmatic function. This observation led to the suggestion that the function of a construction may take precedence over exact frame match.

[1] Czulo uses the term "constructions-and-frames analysis" in Czulo (2013) but since frame comparison is a crucial step in this method (cf. Section 3.2), I will use the term "frames-and-constructions analysis" in this paper.

Building on Czulo's (2013) work, Ohara (2019) and Ohara & Okubo (2020) examined whether frames-and-constructions analysis is a valid methodology to search for and align comparable constructions between Japanese and English, a pair of typologically unrelated languages. That work analyzed 674 pairs of Japanese and English sentences in a bilingual children's book. They identified the semantic frames evoked by the heads of source and target sentences and found 483 pairs of frame mismatches. Among them, 106 pairs exhibited structural divergences. Among the 106 pairs of structurally divergent sentences, 55 pairs exhibited the same pragmatic functions (cf. Table 1, Section 3.3). In other words, the study found corresponding constructions in Japanese and English based on pragmatic functions, even in cases of structural divergence and frame mismatches. Those results suggested the usefulness of frames-and-constructions analysis for finding comparable constructions across typologically unrelated languages such as Japanese and English, where structural divergence is well-documented.

However, the study that Ohara (2019) and Ohara and Okubo (2020) reported is preliminary; and no study exists that explored the validity of the method in analyzing translation from English to Japanese, spoken language, genres other than narratives, and anything other than children's language. Thus, this paper applies the method to analyze English and Japanese sentences that appear in the "Do schools kill creativity?" TED Talk parallel corpus. Analysts already have annotated this corpus with semantic frames and FEs in various languages for Multilingual FrameNet.

3. Frames-and-Constructions Analysis of TED Talk Parallel Transcripts

This section is divided into three parts that describe the following: (1) hypotheses formed prior to the present analysis; (2) details about the proposed method; and (3) results of the analysis on the TED Talk bilingual transcripts.

3.1 Hypotheses

Prior to the present analysis, we formed three hypotheses about characteristics of the English and Japanese sentences in the TED Talk parallel transcripts. First, the TED Talk "Do schools kill creativity?" is a presentation aimed at persuading its audience. Thus, one hypothesis is that the English original transcript would include many constructions that exhibit pragmatic functions. Second, the Japanese version is a translated version of the English original transcript. Consequently, another hypothesis is that the Japanese translation would contain constructions that exhibit similar pragmatic functions as those in the English original. Finally, a third hypothesis is the likelihood of finding corresponding English and Japanese constructions sharing the same or similar pragmatic functions, in spite of also showing frame mismatch and structural divergence.

3.2 Method

The actual adopted steps of the frames-and-constructions analysis in this study appear below. The analysis concentrated on sentence-level grammatical constructions.

1. Head Identification:
   Identify the head of each of the English and Japanese sentence pairs.
2. Frame Comparison:
   Determine the semantic frames evoked by the heads of the English and Japanese sentence pairs; check for frame mismatch; exclude two kinds of cases from frame mismatch. One kind has to do with cases in which a pair of English and Japanese sentences ultimately evokes the same set of semantic frames through frame integration (integration of frames evoked by words and phrases in a sentence that ultimately leads to an understanding of the whole sentence) within each sentence. The other kind involves cases in which the two frames evoked by the English and Japanese heads are related via any FrameNet frame-to-frame relations (Ruppenhofer et al., 2016).
3. Structural Comparison:
   Identify the structure of each of the English and Japanese sentences; check for English and Japanese structural divergence.
4. Functional Comparison:
   Identify the functions of the English and Japanese constructions.

3.3 Results

We examined 242 English original sentences from the TED Talk. Sometimes one English sentence was translated into Japanese with more than one sentence; at other times, multiple English sentences were translated into one Japanese sentence. We concentrated on analyzing sentence pairs in which the English original sentence is more or less straightforwardly translated into Japanese with one sentence. There were 122 such sentence pairs.

Table 1 summarizes the results of our analysis using the steps described in Section 3.2. The table shows the numbers of sentence pairs that exhibit frame match/mismatch, structural divergence, and functional match in the TED Talk parallel corpus, in comparison with those in a bilingual children's book *Anpanman I* (cf. Section 2).

	TED **(E to J)**	*Anpanman* (J to E)
1) Sentence Pairs	**122**	674
2) Frame Match in 1)	**75**	191
2') Frame Mismatch in 1)	**42**	483
3) Structural Divergence in 2')	**22**	106
4) Functional Match in 3)	**9**	55

Table 1: The numbers of frame match/mismatch, structural divergence, and functional match in TED

There was one sentence pair that ultimately evoke the same set of semantic frames through frame integration. In addition, there were four sentence pairs in which the two frames evoked by the English and Japanese heads are related via a FrameNet frame-to-frame relation (cf. Step 3 above). These are the reasons why the sum of the number of frame match and that of frame mismatch does not equal the total number of sentence pairs in the TED Talk parallel corpus.

Let us describe the results in relation to the three hypotheses in Section 3.1. Our first hypothesis was that the English original transcript would contain many constructions that exhibit pragmatic functions. Indeed, the English version of the talk includes sentence structures that focus either the whole or parts of a sentence, such as pseudo-cleft sentences (1), repetition (2), emphasis (3), and cataphora (4).

(1) Pseudo-cleft:

a. *Actually, what I find is everybody has an interest in education.* (#13)
b. *What we do know is, if you're not prepared to be wrong, you'll never come up with anything original -- if you're not prepared to be wrong*. (#77)

(2) Repetition:

a. *What we do know is, if you're not prepared to be wrong, you'll never come up with anything original -- if you're not prepared to be wrong*. (#77)
b. *Picasso once said this, he said that all children are born artists.*(#84)

(3) Emphasis:

My contention is that creativity now is as important in education as literacy, and we should treat it with the same status. (#43)

(4) Cataphora:

a. *Picasso once said this, he said that all children are born artists.* (#84)
b. *If you were to visit education, as an alien, and say "What's it for, public education?"* (#141)

Second, we expected to find in the Japanese translation constructions with similar pragmatic functions as those of the English original. The results of the analysis refuted that expectation. Except for the translation of (4a), listed below as (4'a), which uses cataphora to emphasize a quote from Picasso, none of the Japanese translations of the aforementioned English sentences (1-4) has structures that focus the whole or a part of the sentence or emphasize the speaker's claims. This situation contrasts with that of the English original sentences.

(1')

a. *jissai daremo ga kyôiku ni kanshin ga arundesu*
actually everybody NOM education DAT interest NOM exist
literal translation[2]. 'Actually, everybody has an interest in education.'
b. (=(2'a))
... machigaeru koto o osoreteitara kesshite
make.mistake thing ACC be.afraid never
dokusôteki na mono nado omoitsuk anai
original thing etc. come.up.with NEG
'... if (you are) afraid of making mistakes, (you) will never come up with anything original.'

(3')

sôzôsei wa shikiji nôryoku to onaji kurai
creativy TOP literacy ability COM same degree
kyôiku ni hitsuyô desu
education DAT necessity COP
'Creativity is as necessary to education as literacy.'

(4')

a. (=(2'b))
*Pikaso wa katsute **kô** îmashita*
Picaso TOP once like.this said
"kodomo wa mina umarenagara no âtisuto da"
children TOP all born GEN artist COP
'Picasso once said like this, "children are all born artists."'
b. *moshi eirian ga kyôiku genba ni yatteki tara*
if alien NOM education site LOC come COND
"kô kyôiku tte nan no tame ni aru no?"
public education CONJ what NOM purpose DAT exist Q
to hushigini omou deshô
QUOTE mysteriously think would
'If an alien comes to (an) education site, (s/he) would wonder, "for what purpose does public education exist?"'

The third hypothesis concerned finding English and Japanese constructions that exhibit a structural divergence and frame mismatch, yet have the same pragmatic function. The analysis indeed found instances of such cases. (5) is an example. The heads of the English and Japanese sentences in (5) are *stop* and *surunja arimasen* 'don't!' respectively (Step 1, Section 3.1). The English and Japanese structures are of the `Imperative` construction (cxn) and of `V-surunjanai` cxn respectively (Step 2). The head *stop* in the English sentence evokes the `Activity_stop` frame, while *surunja arimasen* in the Japanese sentence evokes the `Preventing_or_letting` frame (Step 3). Finally, both sentences function to order the addressee to stop an activity (Step 4).

(5) Structural divergence, frame mismatch, and same pragmatic function:

E: *And **stop***[Activity_stop] *speaking like that.* (#105)
J: *sonna hanashi kata **surunja arimasen***[Preventing_or_letting]
that.way speech way don't
'Don't speak like that.'

E: `Imperative` construction (cxn)
J: `V-surunjanai` cxn

E&J: Prohibiting function

4. Discussion

This section discusses functional mismatches in English and Japanese in the parallel transcripts, the validity of the frames-and-constructions analysis method, and the appropriateness of using the TED Talk transcripts for aligning constructions for building a multilingual constructicon.

4.1 Functional mismatches in the TED parallel transcripts

This subsection discusses the results with respect to the second hypothesis in Section 3.1. The second hypothesis in Section 3.1 was that the Japanese translation would

[2] All the translations of the Japanese sentences into English in this paper are literal translations.

contain constructions that exhibit similar pragmatic functions as those in the English original. It turned out that English sentence structures that focus certain of their elements were often NOT translated into Japanese using constructions with similar pragmatic functions.

It may be a consequence of properties of the genre, specifically, of the Japanese transcript. While the English version is an actual transcript of the oral presentation, the Japanese version is primarily a set of subtitles, that is, captions displayed at the bottom of a screen that translate the English transcript. In fact, the sentences in the Japanese version tend to be short and telegraphic, presumably because of the limited space allocated for subtitles and the requirement to be displayed in synch with the audio-visual information in the video clip. Thus, what makes sense is to think of the Japanese transcript as a set of subtitles, something that should be seen and read together with the video clip as part of multimodal information, NOT as a translation. This study has yet to conduct a thorough analysis of the video clip. Some sort of substitute for the pragmatic function to focus a sentence element missing in many of the Japanese sentence structures may be found in the audio-visual information (including speech and gestural information) in the video clip.

4.2 Validity of the Frames-and-Constructions Analysis

Since we were able to found pairs of corresponding constructions in English and Japanese in the TED transcripts, the four steps of the frames-and-constructions analysis proposed in Section 3.2 seem useful in analyzing transcripts of spoken data, in addition to written data. This assessment is legitimate since the concepts embodied in the four steps (i.e., head, sentence structure, semantic frame, and function) are also found in transcripts of spoken data. The proposed four steps particularly emphasize the notion of head. Since the concept is considered universal and since heads can be found in sentences in transcripts of spoken data as well, identifying sentential heads first facilitates accurate linguistic analysis of sentence structures (cf. Croft, In Preparation; Croft et al. 2017).

In this respect, note Lyngfelt et al.'s (2018) proposal concerning alignment of constructions across languages. Based on the analyses of English, Swedish, and Brazilian Portuguese constructions, that work proposed a four-step comparison of constructions (Lyngfelt et al. 2018: 267). The first step is to ask the question "is there a corresponding construction, or set of constructions, in the target language?". While finding corresponding constructions among typologically related languages such as the three languages above may be easy, at least in the case of Japanese and English, identifying corresponding structures is quite difficult. Analyzing a parallel corpus using the frames-and-constructions analysis method, which primarily relies on the concept of head, seems to be a more straightforward way of conducting the analysis.

The proposed four steps of frames-and-constructions analysis predicts that even when frame mismatch and structural divergence are present, if functions are the same, then the two constructions can be considered corresponding. Pairs of constructions exist in the TED Talk parallel transcripts that share the same pragmatic function while exhibiting frame mismatch and structural divergence. It may thus be possible to hypothesize that the function of a construction takes precedence over exact frame match as Czulo suggests. However, it is beyond the scope of this present paper to test this hypothesis.

4.3 Toward a multilingual constructicon

While the TED Talk parallel corpus is a good resource for frame-semantic annotation in individual languages, it may not be the ideal resource as a starting point to align constructions for building a bilingual constructicon between English and another language, because of the characteristics of the genre of subtitles discussed in Section 4.1. We may indeed be able to find better functional alignment between two translated subtitle transcripts, as opposed to comparing one translation to the original[3]. Applying the frames-and-constructions method to translated subtitles of two or more languages may therefore be a better strategy to build a multilingual constructicon from the parallel corpus.

5. Summary and Future Work

This section summarizes the findings of the work presented here:

- Pairs of constructions in English and Japanese that share similar pragmatic functions exist in the TED Talk bilingual transcripts. This is similar to our findings involving texts in a Japanese to English bilingual children's book. Therefore, the proposed frames-and-constructions analysis method seems valid not only for written language but also for transcripts of spoken data.
- While the TED Talk parallel corpus is a good resource for frame semantic annotation in individual languages, it may not be the ideal place to start aligning constructions across typologically unrelated languages, because of the characteristics of the genre of subtitles.
- The frames-and-constructions analysis method proposed here, namely, the one that focuses on the head of a sentence in each language, seems valid to search for corresponding constructions in typologically-unrelated languages.

As Section 4.2 indicates, the four steps of the present frames-and-constructions analysis predicts that even in the case of frame mismatch, and even when structural divergence exists, if the functions of two constructions in the two different languages are the same, then the two constructions are comparable. Croft (In Preparation) and Croft et al. (2017) argue that syntax is primarily motivated by information packaging, and secondarily by semantics. Therefore how the proposed frames-and-constructions analysis method relates to Croft's claim is worth

[3] I would like to thank one of the reviewers for pointing this out to me.

investigating in detail. Of particular interest is how what we have called "pragmatic functions" interacts with Croft's "information packing."

Bibliographical References

Croft, W. (In Preparation). *Morphosyntax: Constructions of the World's Languages*.

Croft, W., Nordquist, D., Looney, K., and Regan, M. (2017). Linguistic typology meets universal dependencies. In Markus Dickinson, et al. editors, Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15), pages 63-75. CEUR Workshop Proceedings.

Czulo, O. (2013). Constructions-and-frames analysis of translations: the interplay of syntax and semantics in translations between English and German. Constructions and Frames, 5(2): 143-167.

Czulo, O. (2017). Aspects of a primacy of frame model of translation. In S. Hansen-Schirra, O. Czulo and S. Hofmann (Eds.), *Empirical modelling of translation and interpreting*. Berlin: Language Science Press, pp. 465-490.

Fillmore, C. J. (2013). Berkeley Construction Grammar. In: T. Hoffmann, & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, pp. 111–132.

Fillmore, C. J. & Baker, C. F. (2010). A frames approach to semantic analysis. In: B. Heine and H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, pp. 313–339.

Lyngfelt, B., Torrent, T.T., Laviola, A., Bäckström, L., Hannesdóttir, A.H., and Matos, E.E.S. (2018). Aligning constructions across languages: a trilingual comparison between English, Swedish, and Brazilian Portuguese. In B. Lyngfelt, L. Borin, K. Ohara, and T.T. Torrent (Eds.), *Constructicography: Constructicon Development across Languages*. Amsterdam: John Benjamins Publishing, pp. 255-302.

Ohara, K. (2019). Frames-and-constructions analyses of Japanese and English Bilingual Children's Book. Paper presented at the Theme Session "Cross-theoretical Perspectives on Frame-based Lexical and Constructional Analyses: Bridging Qualitative and Quantitative Studies" the 15th International Cognitive Linguistics Conference (ICLC 15), Nishinomiya, Japan, August 9.

Ohara, K. and Okubo, Y. (2020). Finding corresponding constructions in a Japanese-English Contrastive children's book using the frames-and-constructions analysis. (In Japanese). Proceedings of NLP2020, pages 921-924, Ibaraki, Japan, March. The Association for Natural Language Processing.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., & Scheffczyk, J. (2016). FrameNet II: Extended theory and practice. Technical report. Berkeley: ICSI.

Language Resource References

Japanese FrameNet.
http://jfn.st.hc.keio.ac.jp

FrameNet.
https://framenet.icsi.berkeley.edu

"Do schools kill creativity?" TED2006, February 2006.
https://www.ted.com/talks/sir_ken_robinson_do_schools_kill_creativity?language=en

Takashi Y. (1991). *Anpanman 1*. Translated by Yuriko Tamaki. Tokyo: Froebel Kan.