**Assignment 5**

**1 Data and Construct**

To persuade the other coder to work for me *without any payment*, I left the choice of research topic to her. The other coder thought the translation of film titles may be interesting. Because when an English film is introduced to China, it often has a translated name which may largely differ from its original English title. Therefore, I collected the related data and we decided to code *the way of film title translation* after discussion.

**1.1 Data pre-processing**

The unit for this text analysis is the dyad of film titles, where each dyad consists of one original title of a European or American film, and its translated Chinese title. However, assuming all raters are English users, these Chinese titles have been back translated to English in a literal way (word-to-word translation). To construct this data set, I took the following steps:

First, I downloaded the [TMDB 5000 data set](https://www.kaggle.com/code/kerneler/starter-tmdb-5000-movies-d190dc26-3) from Kaggle.com.

Second, I chose the Top 1000 popular films through sorting the data by the films’ vote counts (the number of people who voted for this film), which I think reflect the popularity of films.

Third, I used GPT to find the officially translated Chinese names for these films.

Fourth, after deleting names which is obviously wrong, I used GPT again to translate this Chinese name back to English with asking the GPT to translate in a literal way.

Then, I used the python code of *Levenshtein Ratio* to calculate an initial text similarity. I dropped samples whose *Levenshtein Ratio <* .5, in order to get a subsample *with large variance*. This results in a set of 328 samples.

At last, I resort the sample with vote counts and retain the most popular 50 films. This is the sample we used in this assignment.

**1.2 Construct Decision**

We firstly wanted to code a simple construct and a more complexed latent construct. For the simple construct, we chose to find out the extent to which the translated title is similar to the original one, or putting it formally but differently, the level of free translation (LFT). However, we soon found that this construct is not as simple or direct as we thought after we meet lots of challenges which would be discussed later.

For the latent construct, we had planned to code the change of 5W1H elements (Who, What, When, Where, Why, How) between original title and the translated title. We soon realized the labor for this work would be too tough. Meanwhile, follow suggestion (Neuendorf, 2017) that raters should not be asked to make multiple judgement for one construct, we simplified this construct to only comparing the who-class object between the two titles, i.e., to put it formally, whether there is additional explicitation of actors in the translated title (AEATT). That is, whether the translated title wants to better highlight the who-class objects in the film.

**2. Process Design and Decision**

**2.1 Process**

Briefly, I split the 50-sample data set into three subsets (15+15+20). Before coding any subset, we discuss to the way to set or modify the codebook. Then, the two coders code independently. After coding any subset, Cohen’s kappa is calculated and discussion is initiated until we think that consensus is reached with the modification of the codebook. For the three rounds of coding, below is the Cohen’s kappa statics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Round | Sample size | Agreement for LFT (Expected agreement) | Cohen’s kappa for LFT (Prob >Z) | Agreement for AEATT (Expected agreement) | Cohen’s kappa for AEATT (Prob >Z) |
| 1 | 15 dyads | 46% (21%) | .32 (.01) | 80% (40%) | .67 (.00) |
| 2 | 15 dyads | 73% (38%) | .57 (.00) | 87% (57%) | .68 (.00) |
| 3 | 20 dyads | 80%  (31%) | .71 (.00) | 90% (75%) | .61 (.00)\* |

\* It can be noticed here that after iterative improvement of coding, the Cohen’s kappa for AEATT in the 3rd round **decreases unexpectedly!** This seems not reasonable. Actually, I think this is the issue of sample. AEATT is a four-scale measurement, but as for the last 20 sample, only two values (0 and 1) are observed. Meanwhile, I don’t know how to tell the Stata that AEATT should theoretically have four different values.

**2.2 Decisions**

The first challenge is, when coding LFT, raters tend to measure the similarity of the two titles by building connection with information from the content of the film. This soon raises problems because raters may not have watched each film. Low agreement raises between raters because they have difference watching experience. Therefore, we find it is necessary to assume rater has *no any knowledge* from the content of the film. This problem doesn’t only show in round 1, but also shows again in round 2 after we have reached this agreement, because raters may remember the content of the film (not deliberately) and then are allured to make bias judgement. Therefore, we highlight in the codebook as strong as possible to ask the rater to constrain themselves not to drawn on knowledge of the film content. They must make judgement only based on the information offered by the two titles.

The second great challenge is for LFT ratings. In the beginning, LFT has 5 levels of ratings. Now in the final codebook, there are only four. The fifth rating raise unsolvable inconsistence. Therefore, after trying, we deleted this rating. The fifth rating is ‘*4 – The two titles should be viewed as totally irrelevant, i.e. the rater struggles to establish a connection between the two titles, even when compelled to do so*’. We wanted to draw a line between ‘4’ and ‘3’ (*3 – There are no any similar elements in the two title or all elements are identified different, but the rater can construct weak connection between the two titles through contextual imagination*) but we failed. Coders often disagree on what is the condition for ‘*totally irrelevant*’. When one rater finds it ‘*totally irrelevant*’, the other may somehow manage to build a weak relevance between the two titles by really remote imagination.

The third challenge (or a group of challenges) concerns the definition of words we used in item illustration. For example, what is the meaning of ‘ignorable changes’, ‘synonyms’, ‘elements’, ‘similar elements’ and ‘actor/character’. Discussion for this challenge results in many times of modification on the codebook and additional attention reminders for many items.

**3. Future Memo**

For future continuation, I think the coders must be well trained and be tested after training before they engage in the formal process of coding. This is because the task I have done reveals that this coding requires coders to resist their natural tendency to recall the content of films. They have to fully invoke their rationality to assume they hadn’t known something that they do know.

Besides, I will further decompose the AEATT construct, i.e., not to ask raters to compare between original title and translated title, but to ask raters to simply count the number of who-class objects in all kinds of title (we don’t tell raters whether this title is an English original title or a Chinese back translated title). Then we measure AEATT for each title dyad by calculating the difference between the respective count of who-class objects in the two titles.

**Appendix**

**Codebook**

**Unit of Coding:**

**The dyads of film titles.** Each dyad consists of one *original title of a European or American film*, and *its translated Chinese title*. However, assuming all rater are English users, these Chinese titles have been back translated to English in a *literal* way (word-to-word translation).

**Instruction and Qualification of Raters:**

First, raters should be a good commander of English language, i.e., raters should know of some basic usages or habits in English vocabulary and gramma. When raters meet some words that he or she doesn’t know or feels not sure, she *can and should* search it on the Internet. The rater should be able to understand the word *only through* reading the English explanation of the word rather than explanation in other languages, such as Chinese.

Second, raters do not need the knowledge of the content of the films he or she is coding. More important, in the whole process of coding, the rater should assume that he or she doesn’t know the content of the film. If the raters indeed have watched the film, they should constrain themselves *not to* make judgement by referring to the information beyond the title.

Third, raters should first read this codebook thoroughly! Pay attention to all item illustration below, in particular, the explanation marked with ‘**Attention here**’.

**Construct 1: Level of Free Translation**

By comparing the original title and the translated title, the rater should judge to what extent this title has been translated in the *free style* rather than a *literal style*. Free style translation refers to that translator paraphrases the original title, which results in a significant change in naming the film. Literal style translation refers to that the translator keeps loyal to the original title when translating. Putting it in a different way, this construct is asking the rater to give a rating which can reveal how the translated title is deviated from the original one. The rating should fall in the interval from 0 to 3. ‘0’ means the translated title is almost the same as the original one, and ‘3’ means that translated title are totally different. Each number is illustrated in detail below:

0 – Word-to-word translation with almost no change on the use of words, or with ignorable changes, such as the morphological change or derivation of words, e.g., from *Zombieland* to *Land of Zombies(僵尸之地).*

1 – Connection between the two titles can be easily inferred from the similarity between the meaning of the words they use or, to put it simply, from the use of synonyms, e.g., from *Gone Girl* to *The Disappeared Girl.*

**Attention here:** Synonyms for character names can refer to different names for the same character in the English culture, e.g., Thor and the Thunder God.

2 – Even though the translated title consists of new or different elements, but the connection between the two titles can still be revealed by *remaining similar elements*, e.g., from *The Jungle Book* to *Forest Wonder(森林奇缘).*

**Attention here:** Elements are defined by words carrying informative content. For instance, prepositions can’t be viewed as elements. It is also important for raters to realize that, *similar elements* can be synonyms but *don’t limit to synonyms*. This implies that the condition for the rating of ‘1’ is more narrowed and specific than that of the rating of ‘2’. For example, the word *sea* and *ocean* can be view as both synonyms and similar elements, but the word *sea* and *water* can only be view as similar elements, not synonyms. The difference of the judgment between the rating of ‘1’ and ‘2’, is that ‘2’ does not only loose the requirement from synonyms to similar elements, but also allows for different elements being added into the title.

3. All elements in the two titles are identified different. Different elements might be somehow interrelated, but this relationship demands rater’s active imagination, e.g., from *Die Hard* to *Impossible Mission(不可能的任务)*.

**Construct 2: Additional Explicitation of Actors in Translated Title**

Additional explicitation of actors in translated title refers to whether the Chinese translator of the title adds additional objects into the title which maybe the potential character or actor of the film. However, this construct *doesn’t care about* whether the object added is or is not the real actor in the film. This construct is asking the rater to compare the English title and the Chinese title on which title strength more on concrete referents rather than other elements of film, such as time, location, or relationship. This ‘referent’ is defined here as who-class objects, like the actor, people, character, or other animate referents. Follow the instruction bellow to make this comparison:

-1 - Original title mentions the actor, people, character, or other animate referents (i.e., who-class objects), but the translated title doesn’t

0 - The original title and the translated title both mention or both not mention the actor, people, character, or other animate referents (i.e., who-class objects)

1 - Original title doesn’t mention the actor, people, character, or other animate referents (i.e., who-class objects), but the translated title does

99 - Not Clear

**Attention here:**

First, animate referents, i.e., who-class objects can be objects more than just human beings. It is quite usual for animals, robots, warms, ghosts or even cars, toys and vegetables to be the characters of films. The raters should refer to the context of the title to make appropriate judgement.

Second It is not uncommon that the rater may see a name of something in the title: Sometimes, the rater can infer that the name is referring to a person or a role through the language habits of English, for example, Mr. <suppose a name here>. However, there is also time when the rater can not infer what the name is for, for example, a film title can be *NX.* This name may or may not be a name for a who-class objects. It may be a where-class object, e.g., a name for the location in the story, which, if it is in this can, is not of the interest of our research. When the rater is not sure whether a name is referring to a who-class object (after considering the language habits and the context information of the title), then he or she should reasonably choose ’99 - Not Clear’. What is more important here, since the rater should assume that he or she has no any knowledge of the content of the film in coding, he or she should not use the information from the content of the film (if she knows of it) to infer whether the name is for a who-class object.

**Code Form**

* LFT = Level of free translation
* AEATT = Additional explicitation of actor in the translated title

|  |  |  |  |
| --- | --- | --- | --- |
| **Original Title** | **Chinese Translated Title (After Literal Back Translation to English)** | LFT  [0,1,2,3] | AEATT [-1, 0, 1, 99] |
| Inception | Dream Space Thief |  |  |
| Django Unchained | The Liberated Django |  |  |
| Guardians of the Galaxy | Galaxy Guardians |  |  |
| The Shawshank Redemption | Moon Black High Fly |  |  |
| Skyfall | 007: The Broken Sky Assassination Machine |  |  |
| Batman Begins | Batman: The Mystery of the Heroic Shadow |  |  |
| Up | Flying House Travelogue |  |  |
| Inside Out | Mind Agents Team |  |  |
| Thor | Thunder God |  |  |
| Inglourious Basterds | Shameless Bastard |  |  |
| Man of Steel | Iron Man |  |  |
| Shutter Island | Mysterious Shadow Infatuation |  |  |
| WALL·E | Robot Mobilization |  |  |
| Big Hero 6 | Super Power Land Battle Team |  |  |
| Finding Nemo | Underwater Mobilization |  |  |
| Gone Girl | The Disappeared Girl |  |  |
| Life of Pi | The Fantastic Drift of Pi's Youth |  |  |
| Now You See Me | Multiple Shadows |  |  |
| Toy Story | Toy Mobilization |  |  |
| Oblivion | Forgotten Planet |  |  |
| Silver Linings Playbook | The Clues of Happiness Behind the Dark Clouds |  |  |
| Ex Machina | Mechanical Maiden |  |  |
| Looper | Future Police |  |  |
| Wreck-It Ralph | Invincible Wrecking King |  |  |
| Maleficent | Sleeping Curse |  |  |
| Spectre | 007: Ghost Party |  |  |
| V for Vendetta | Assassin's Creed |  |  |
| The Prestige | Fatal Magic |  |  |
| Ratatouille | Food Total Mobilization |  |  |
| Taken | Special Forces |  |  |
| The Departed | Infernal Affairs |  |  |
| Whiplash | Exploding Drummer |  |  |
| How to Train Your Dragon | Dragon Tamer |  |  |
| Furious 7 | Speed and Passion 7 |  |  |
| Her | She |  |  |
| Die Hard | Impossible Mission |  |  |
| Fury | Wolf Warrior Action |  |  |
| Cars | Car Racing Total Mobilization |  |  |
| Insurgent | Rebel. |  |  |
| Catch Me If You Can | Incredible Thief |  |  |
| Ocean's Eleven | Eleven Arhats |  |  |
| The Fault in Our Stars | Our Star's Tragedy |  |  |
| Reservoir Dogs | Bad Dog Gang |  |  |
| Eternal Sunshine of the Spotless Mind | The Disappeared Light Years |  |  |
| The Bourne Identity | Multiple Shadows |  |  |
| Zombieland | Land of Zombies |  |  |
| Blade Runner | Silver Wing Killer |  |  |
| X2 | X-Men 2 |  |  |
| A Good Day to Die Hard | Need for Speed 6: Wild Speed |  |  |
| The Fast and the Furious | Speed and Passion |  |  |