Opera Goer or Scrabble Player - What Makes a Good Translator?

Naoto Nishio
Department of CSIS
University of Limerick, Ireland
naoto.nishio@ul.ie

Richard F.E. Sutcliffe
School of CSEE
University of Essex, UK
rsutcl@essex.ac.uk

Abstract

How should you select a person to carry out a translation? One approach is to request a sample translation and to evaluate it by hand. Quality Estimation (QE) addresses the problem of evaluation at least for Machine Translation output as a prediction task (Specia et al. 2010). This approach facilitates low-cost evaluation of MT outputs without expensive reference translations. However, the prediction of human translation in this way is difficult due to its subtlety of expression (Specia and Shah 2014). We aimed to find out whether the qualifications, hobbies or personality traits of a person could predict their proficiency at translation. First, we gathered information about 82 participants; for each one we established the values of 146 attributes via a questionnaire. Second, we asked them to carry out some Japanese-to-English translations which we scored by hand. Third, we used the attributes as input and the translation scores as output to train the J48 decision-tree algorithm in order to predict the score of a translator from their attributes. This was then evaluated using ten-fold cross validation. When limiting to professional translators in Experiment 6, the best F-score was with Wrapper selection (0.775). The result was statistically significant ($p < 0.05$). This classifier also showed the second highest Precision on Good (0.813). The second best F-score (0.737) has the highest Precision on Good (0.909), using Manual feature selection. Once again this was significant ($p < 0.05$). The results suggest that certain attributes affect the prediction; in addition to experience and qualifications in translating into the target language, interest in going to the Opera, playing Scrabble or Contract Bridge, or enjoyment of cryptic crossword puzzles are important factors as well.

Keywords: evaluation of human translation, manual evaluation of translation quality, prediction of translation score, machine learning, decision tree, features of a translator

1. Introduction

In a multilingual world there is an ever-increasing need for translation, some of which must be carried out by human translators. When a translator is to be chosen by hand, qualifications, work experience and personal references are often consulted (Schopp 2007; Biel 2011). Furthermore, sample translations are typically requested and evaluated before a final selection decision is taken. However, the manual evaluation of such translations is a difficult and costly task which depends on skilled evaluators who understand the content - both in the source and target language.

One alternative is to use automated evaluation metrics such as BLEU (Papineni et al. 2002), WER (Popović and Ney 2007) and TER (Snover et al. 2006). However, they require reference translations to be created and still they only approximate to human judgment. Indeed, Turian et al. (2003) observe that the correlation between human judgements and automatic evaluation measures is low. Quality Estimation (QE) avoids the need to create reference translations by using machine learning applied to various features of a sentence and its translations (Avramidis et al. 2011; Blatz et al. 2004; Specia et al 2009; Specia et al 2010). QE has been extended to the evaluation of translations produced by

human translators (Specia and Shah 2014). In another line of research, reading comprehension and summarisation show positive indications of translation ability (Tavakoli et al. 2012; Rahemi et al. 2013; Brau 2014).

However, is there a way to predict translation ability without setting any translation task for a candidate to perform? Answering this question was the aim of this study. The starting point was the informal observation that some types of person seem more suited to being translators than others. Aside from obvious factors such as experience and qualifications, we wanted to investigate the effect of a person's interests, for example in music or literature, their hobbies, such as playing board games or reading the newspaper, and finally their personality traits, such as reversing their car into the driveway ready to leave easily the next morning.

Accordingly, we started by recruiting 82 participants, each of whom filled in a questionnaire. This enabled us to find their qualifications, interests, hobbies, and traits, which were expressed as a set of 146 attributes. Next, the participants carried out some translations from Japanese to English, using complex texts taken from the proceedings of the Japanese parliament. These were evaluated by hand using a Nugget Recall and Fluency (NRF) metric developed by us.

Third, now that we had a measure of how good the participants were at translation, we attempted to predict this from their attributes using machine learning. We used a translator's attributes as input and their translation score as the required output to train the J48 decision-tree algorithm in order to predict a translator's ability automatically from their attributes.

In the rest of this article we start by discussing related work in Section 2. In Section 3, we describe the attributes used for the machine learning and how they were elicited by a questionnaire. Sections 4 and 5 deal with the selection of texts for the work and the method by which evaluation of translations by participants was carried out. Section 6 describes how participants were recruited and what their characteristics were. Section 7 describes six experiments which were carried out on the resulting data. Finally, the last section draws conclusions from the work.

2. Background

Concerning the prediction of translation quality from the characteristics of a translator, natural criteria are their qualifications and experience. For example, Gile (1995) argues that formal training enhances a translator's performance, while Biel (2011) suggests that a certified translator is capable of high quality work. European standard EN 15038:2006 states that a translator should have a degree in translation, subject matter qualification and years of experience (Schopp 2007; Biel 2011). Living in the target language country for a while can also be an important factor (Napthine 1983).

What about skills, interests or personality traits? Napthine (1983) advocates a willingness to track down information and a well developed critical sense; Verrinder (1983) mentions familiarity with current affairs and a temperament suited to working alone; The European Commission (2015) makes reference to a multicultural mind set and a grasp of complex issues; Nida (1981) emphasises the need to have a creative imagination and to be capable of using words imaginatively; Suzuki (1988), based on a survey of translators, mentions an interest in the arts, especially in writing novels, screen plays, drama, haiku, and poems, and also an interest in intercultural aspects of daily life.

Hubscher-Davidson (2009) considers the effect of personal traits on the quality of translation. The author uses seven categories: Introversion, Extraversion, Sensing, Intuition, Thinking, Feeling, Judging, and Perceiving referring to the well-known Myers-Briggs Type Indicator (MBTI) (Myers & Myers 1995). The findings suggests that Sensing and Intuition are important to a good translator. O'Brien (2013) specifically refers to the relationship between personality and the translation process and mentions that there has been little work on it other than Hubscher-Davidson (2009).

Based on the above work, we track qualifications, experience, residence, interests and various personality traits in our study in order to establish how important they are. See Section 4 below for more details.

Turning to measures for evaluating translators based on their work, a common approach is to use Adequacy and Fluency (Koehn 2010). The use of a scale of one to five for each is a popular approach for machine translation development (Callison-Burch et al. 2007; Koehn 2004; Sumita et al. 2005) and this idea dates back to Nagao et al. (1985). At the WMT evaluations, a manual method based on ranking has been used (Bojar et al. 2014). Human evaluators (mostly WMT participants) are shown a source sentence, a reference translation, and five candidate translations. They then rank the candidates from best to worst. Each participant carries out 300 such tasks. By combining the results over many systems and many test sentences it is possible to rank all the systems. The TrueSkill algorithm for combining ranks (Sakaguchi et al. 2014) was found to have the highest correlation with human judgements. However, Graham et al. (2015) criticise 'notoriously inconsistent' human assessment in the WMT ranking task and investigate the use of a single continuous Likert scale for translation evaluation, along with a large number of evaluators recruited via Mechanical Turk. They found that by taking the mean score of fifteen assessors for each sentence to be evaluated, it was possible to obtain a Pearson correlation of 0.9 with the mean scores computed from a very large population of assessors. When the number of assessors rose to 40, the correlation reached 0.97.

For our own work, we have developed a method called Nugget Recall and Fluency which is based on the Nugget Recall of Vorhees (2004). As we will see in Section 6.2, this tests whether or not a candidate translation contains the key points of the reference rather than relying on the general notion of Adequacy. It is therefore suitable for very complex translations testing the ability of advanced translators.

Concerning automatic measures, the BLEU algorithm of Papineni et al. (2002) has had enormous influence. However, to work best it requires not one but several reference translations per sentence and it should also be applied to a large number of test sentences to get an accurate result. The popularity of BLEU and the previous existence of PER (Tillmann et al. 1997) has led to many derived measures including CDER (Leusch et al. 2006), METEOR (Banerjee and Lavie 2005), NIST (Doddington 2002), TER (Snover et al. 2006) and WER (Popović and Ney 2007). Each tries to overcome different limitations of a method based fundamentally on automatic comparisons which do not assess intrinsic translation quality *per se*.

Interesting as these measures are, they all require reference translations with which to compare a candidate translation for a particular source sentence. On the other hand, Quality Estimation (QE) predicts the quality of an MT system output from a set of inputs without any information about the expected output (Specia et al. 2009). QE has been actively studied for speech recognition but it has not been well-known in other areas of NLP such as MT (Blatz et al. 2004). Specia et al. (2009) and Specia et al. (2010) take QE as a regression problem of continuous translation quality scores between features (from the translation process, input sentences, and translation sentences) and their quality scores. While QE may be a promising approach for the evaluation of hypotheses, QE requires a large amount of training data to carry out prediction because the estimation is based mainly on a set of features of sentences. QE is therefore actively used for MT evaluation tasks, which can generate a large amount of data to analyse the improvement of algorithms. Specia and Shah (2014) explore the prediction of quality of translations produced by professional human translators based on features of source and target sentences. They conclude that in general predicting human translation quality is harder because HT errors may be more subtle and require more sophisticated features than currently used ones. Zaidan and Callison-Burch (2011) explore a method to select high quality professional equivalent translations from sentences translated by inexpensive non-professional translators as a prediction problem, which is set as a binary question between acceptable and not acceptable based on a model of features in three groups: sentence-level, worker-level (the native language, the duration of use of languages, locations, etc.), and ranking. It suggests that casual translators can translate as well as professional ones. Yet, inconsistency of human evaluation is little addressed. Moreover, the different features between a professional translator and a casual one are not clear apart from their price difference.

In MT automated evaluation metric development, the human assessment is the gold standard for evaluating the development against its predecessors (Graham et al 2015). Little consideration is given to the selection of human assessors, evaluators or translators except for their hiring cost and the

amount of time the manual task requires (Zaidan and Callison-Burch 2011; Bojar et al. 2013; Graham et al 2015).

3. Participants

Having first sought and received approval from the University of Limerick Ethics Committee, we identified the following as sources of participants: Associations of English and Japanese teachers, associations of translators in Japan and countries that used English as a common language, online networking sites for translators, translation service providers, language teaching institutions, and language courses in Universities. After extensive recruitment, a total of 82 participated in our study. The participants were not limited to professional translators. While a participant could not be a minor (i.e. under eighteen), anyone who understood complex Japanese texts and was able to translate them into English was considered a valid candidate. Furthermore, we did not limit the participation to native speakers of English or Japanese. Approximately 46 percent of the 82 participants were native English speakers and 41 percent were native Japanese speakers. Half of the participants were fluent in English and 28 percent in Japanese.

4. Elicitation Questionnaire

4.1 Outline

We defined six categories of attributes: Arts (literature, music, paintings), Sports, Pastimes, Lifestyle, Personality, and Background. The questionnaire was designed to collect values for the attribute labels under each category from the participants. However, a small number of participants did not return answers for some questions.

4.2 Arts

We envisaged that familiarity with the arts might indicate a good translator; this was based on Nida (1981) who found that creative imagination was important, and Suzuki (1988) who discovered that an interest in the arts was a positive indicator. Questions were concerned with English literature – both novels and plays (Table 1), Japanese literature (Table 2), classical music (Table 3), and paintings (Table 4). In each case we aimed to choose works which were varied and well-known.

English novel		**English poem**	
To the Lighthouse - Virginia Woolf		Shall I compare thee to a summer's day - William Shakespeare	
Brideshead Revisited - Evelyn Waugh		You are old Father William - Lewis Carroll	
Bleak House - Charles Dickens		Tyger Tyger, Burning Bright - William Blake	
None		None	
Familiar with three	5	Familiar with three	5
Familiar with two	4	Familiar with two	13
Familiar with one	10	Familiar with one	22
Not familiar with any	63	Not familiar with any	42
Preference		Preference	
To the Lighthouse - Woolf	6	Shall I compare thee to a summer's day - Shakespeare	23
Brideshead Revisited - Waugh	5	You are old Father William - Carroll	4
Bleak House - Dickens	8	Tyger Tyger Burning Bright - Blake	13

Table 1: English Literature Features. The tables in this section indicate what information was stored in features used for ML training. This data was obtained from participants via a questionnaire. The top portion shows three novels and three poems. The middle portion shows how many novels and poems each participant was familiar with. The bottom lists how many preferred a given novel over the other two, and the same for each poem. This results in six Boolean features (one for each work) and two further multi-valued features, one for the preferred novel and one for the preferred poem.

Japanese novel		**Japanese poem**	
吉里吉里人 (Kirikiri jin) Title: Kirikiri People 作家　井上ひさし Author: Hisashi Inoue		雨にも負けず (Ame nimo makezu) Title: I will not lose to the rain 詩人　宮沢賢治 Poet: Kenji Miyazawa	
三四郎 (Sanshirō) Title: Sanshirō (name of a character in the book) 作家　夏目漱石 Author: Sōseki Natsume		君死にたまふことなかれ (Kimi shinitamou koto nakare) Title: You had never died 詩人　与謝野晶子 Poet: Akiko Yosano	
蟹工船 (Kani kō sen) Title: Crab Fishing Vessel 作家　小林多喜二 Author: Takiji Kobayashi		こころ (Kokoro) Title: Heart 詩人　萩原朔太郎 Poet: Sakutarō Hagiwara	
None		None	
Familiar with three	3	Familiar with three	10
Familiar with two	12	Familiar with two	13
Familiar with one	16	Familiar with one	21
Not familiar with any	51	Not familiar with any	38
Preference		Preference	
吉里吉里人 – Inoue	5	雨にも負けず – Miyazawa	32
三四郎 – Natsume	16	君死にたまふことなかれ – Yosano	7
蟹工船 – Kobayashi	10	こころ – Hagiwara	5

Table 2: Japanese Literature Features. The top shows three Japanese novels and three poems. The middle indicates how many participants were familiar with them. The bottom shows the preferred novel and poem. This results in six Boolean features (one for each work) and two further multi-valued features, one for the preferred novel and one for the preferred poem.

Works of Classical music		**Japanese songs**	
Title: Magnificat in D major, BWV 243 Composer: Johann Sebastian Bach		「赤とんぼ」(Aka tonbo) Title: Red Dragonfly 三木露風 作詞 山田耕筰 作曲 Poet: Rofū Miki Composer: Kōsaku Yamada	
Title: Symphony No. 5 in C minor, Op. 67 Composer: Ludwig van Beethoven		「早春賦」(Sō shun fu) Title: Early Spring 吉丸一昌 作詞 中 田 章 作曲 Poet: Kazumasa Yoshimaru Composer: Akira Nakata	
Title: Piano Concerto No. 1 in B-flat minor, Op. 23 Composer: Pyotr Ilyich Tchaikovsky		「夏の思い出」(Natsu no omoide) Title: A Memory from Summer 江間章子 作詞 中田喜直 作曲 Poet: Shōko Ema Composer: Yoshinao Nakada	
Title: Élégie for Cello and Piano, Op. 24 Composer: Gabriel Fauré		「花」(Hana) Title: Flower 武島羽衣 作詞 滝廉太郎 作曲 Poet: Hagoromo Takeshima Composer: Rentarō Taki	
Title: A Flock Descends into the Pentagonal Garden Composer: Toru Takemitsu		「浜辺の歌」(Hamabe no uta) Title: Beach Song 林 古 溪 作詞 成田為三 作曲 Poet: Kokei Hayashi Composer: Tamezō Narita	
None		None	
Familiar with five	2	Familiar with five	13
Familiar with four	6	Familiar with four	6
Familiar with three	22	Familiar with three	3
Familiar with two	20	Familiar with two	13
Familiar with one	14	Familiar with one	12
Not familiar with any	18	Not familiar with any	35
Preference		Preference	
Magnificat – Bach	7	「赤とんぼ」 三木露風 作詞 山田耕筰 作曲	23
5th Symphony – Beethoven	34	「早春賦」 吉丸一昌 作詞 中 田 章 作曲	4
Piano Concerto No. 1 – Tchaikovsky	19	「夏の思い出」 江間章子 作詞 中田喜直 作曲	8
Élégie for Cello – Fauré	3	「花」 武島羽衣 作詞 滝廉太郎 作曲	8
A Flock Descends into the Pentagonal Garden - Takemitsu	1	「浜辺の歌」 林 古 溪 作詞 成田為三 作曲	4

Table 3: Classical Music Features. The top shows five works of classical music and five classical Japanese songs. The middle indicates how many participants were familiar with them. The bottom shows the preferred work and song. This results in ten boolean features (one for each work) and two further multi-valued features, one for the preferred piece of classical music and one for the preferred Japanese song.

Paintings	
Fireworks in Nagaoka - Kiyoshi Yamashita	
The Night Watch - Rembrandt van Rijn	
Peasant Wedding - Peter Brueghel the Elder	
The Hay Wain - John Constable	
Fighting Temeraire - J. M. W. Turner	
Bathers at Asnières - Georges Seurat	
None	
Familiarity	
Familiar with six	1
Familiar with five	1
Familiar with four	3
Familiar with three	9
Familiar with two	15
Familiar with one	19
Not familiar with any	34
Preference	
Fireworks in Nagaoka - Kiyoshi Yamashita	6
The Night Watch - Rembrandt van Rijn	24
Peasant Wedding - Peter Brueghel the Elder	6
The Hay Wain - John Constable	2
Fighting Temeraire - J. M. W. Turner	4
Bathers at Asnières - Georges Seurat	6

Table 4: Paintings Features. The top shows six paintings. The middle indicates how many participants were familiar with them. The bottom shows the preferred painting. This results in six Boolean features (one for each painting) and a multi-valued feature for the preferred painting.

4.3 Sports

The European Commission (2015) states that a good translator is durable under pressure, self-disciplined and consistent at work. As such qualities may also be possessed by those who are successful at sport, we decided to ask about this topic. There were two questions: one was concerned with team sports and the other focused on individual sports. Table 5 shows that more participants took part in individual sports than team sports. Perhaps this concurs with the idea that a translator must be good at working on their own (Verrinder 1983).

Team sports		**Individual sports**	
Baseball		Cycling	
Hockey		Ice/roller skating	
Rugby		Running/jogging	
Soccer		Swimming	
		Snowboarding/skiing	
		Familiar with five	5
Familiar with four	0	Familiar with four	0
Familiar with three	5	Familiar with three	11
Familiar with two	5	Familiar with two	12
Familiar with one	13	Familiar with one	29
Not familiar with any	59	Not familiar with any	25

Table 5: Sports Features. The top shows four team sports and five individual sports. The bottom indicates how many participants were familiar with them. This results in eleven Boolean features, one for each of the nine sports, and two further Boolean features, one for no team sport and one for no solo sport.

4.4 Pastime Activities

The four questions in this section were concerned with puzzles, card games, and board games. It was thought that those who played such games might have 'information acquisition ability', 'linguistic knowledge', 'a grasp of complex issues' and ability at 'communication and information management' – important translator qualities identified by the European Commission (2015).

Puzzle		Card game		Board game	
Sudoku		Bridge		Chess	
Crossword puzzle		Hanafuda		Scrabble	
Kanji crossword		Poker		将棋 (Shōgi)	
Jigsaw puzzle		Blackjack		碁 (Go)	
None of these		百人一首 (Hyakunin Isshu)		麻雀 (Mahjong)	
		None of these		None of these	
Familiar with four	1	Familiar with five	1	Familiar with five	1
Familiar with three	9	Familiar with four	3	Familiar with four	1
Familiar with two	19	Familiar with three	6	Familiar with three	2
Familiar with one	25	Familiar with two	19	Familiar with two	20
Not familiar with any	28	Familiar with one	16	Familiar with one	28
		Not familiar with any	37	Not familiar with any	30

Table 6: Pastime Features. The top shows four puzzles, five card games, and five board games. The bottom indicates how many participants were familiar with each of them. This results in eighteen boolean puzzle features, four for puzzles, five each for card games and board games, one each for no puzzle, no card game and no board game, and one for an additional question about cryptic puzzles. Hyakunin Isshu is a Japanese card game based on Japanese classical poems; Shōgi is a Japanese war board game, similar to Chess; Go is a Chinese board game using black and white stones to mark territory; Mahjong is a tile game, similar in concept to the card game Poker, which originated in China.

4.5 Lifestyle

Twenty questions were designed to ask about participants' lifestyle. These dealt with performing music singly or in groups, attending performances of music, theatre etc., reading newspapers, and exhibiting general traits. Suzuki (1988) found that involvement in the performing arts could be significant; concerning newspapers, Verrinder (1983) mentions an interest in current affairs. The general traits are more speculative, though they have some link to the translator personality work undertaken by Hubscher-Davidson (2009).

Musical instrument		Musical ensemble	
Piano	29	Orchestra	6
Violin	4	Choir	3
Guitar	12	Rock/pop band	6
Flute	7		
Recorder	6		
Drum	5		
Other	5		
Play six instruments	0	In four ensembles	0
Play five instruments	0	In three ensembles	0
Play four instruments	0	In two ensembles	2
Play three instruments	5	In one ensemble	11
Play two instruments	12	In no ensemble	69
Play one instrument	29		
Play no instrument	36		

Table 7: Lifestyle - Music Playing Features. The top shows six musical instruments and three music ensembles. The bottom indicates how many instruments participants played and how many ensembles they were members of. This results in twelve Boolean features, seven for instruments including 'Other', three for ensembles, and one each for no musical instrument and no ensemble.

Performing arts	
Concerts (any type)	59
Theatre	35
Operas	14
Films	70
Enjoy four of them	8
Enjoy three of them	30
Enjoy two of them	21
Enjoy one of them	15
None of them	8
Rakugo	
Yes	36
No	46

Table 8: Lifestyle - Performing Arts Features. The top shows four types of performing art. The middle indicates how many the participants attended. The bottom shows how many participants were familiar with Rakugo which is a traditional Japanese performing art. This results in six Boolean features, one for each performing art, one for no performing art, and one for Rakugo.

English newspaper		Japanese newspaper	
The Daily Telegraph	1	Asahi	24
The Financial Times	4	Chunichi	2
The Guardian	16	Mainichi	2
The Irish Independent	6	Nikkei	13
The Irish Times	14	Sankei	1
The Daily Mail	1	Yomiuri	8
The Times	13	Tokyo sports	1
Familiar with seven	0	Familiar with seven	1
Familiar with six	0	Familiar with six	3
Familiar with five	2	Familiar with five	2
Familiar with four	5	Familiar with four	4
Familiar with three	11	Familiar with three	7
Familiar with two	18	Familiar with two	16
Familiar with one	19	Familiar with one	18
Not familiar with any	27	Not familiar with any	31
Familiar with both English and Japanese papers			34
Familiar with either English or Japanese papers, not both			38
Familiar with neither			10

Table 9: Lifestyle - Newspaper Features. The left side shows seven newspapers in English while the right side shows seven in Japanese. The middle shows the number of newspapers in English or Japanese with which the participants were familiar. The bottom shows how many were familiar with papers in both languages or just one. This results in fourteen Boolean features, one for each of the fourteen newspapers, and two multi-valued features, one for the preferred English newspaper and the other for the preferred Japanese newspaper.

	Yes	No	Selection of fruit juice	
Bilingual environment	16	66	Apple	18
TV at home	76	6	Grapefruit	17
Prefer radio to TV	56	26	Orange	35
Fun to read a book	67	15	Pineapple	9
Like to ride a bicycle	70	12	None	3
Primary education at one place	27	55	Strong hand	
Friend from abroad	78	4	Left hand	7
Use of library	74	8	Right hand	74
Mechanical pencil sharpener	38	44	Ambidextrous	1
			Wearing wristwatch	
			Left wrist	33
			Right wrist	6
			No wristwatch	43

Table 10: Lifestyle - General Features. The left side shows nine general lifestyle characteristics. The top right shows preference for different fruit juices. The middle right shows handedness and the bottom right indicates the wristwatch hand. This results in nine Boolean features, one for each lifestyle aspect on the left, and three multi-valued features, one each for preferred fruit juice, strong hand and wristwatch hand.

4.6 Personality - Association, Imagination, and Empathy

There were ten questions regarding wider aspects of personality such as associations linked to concepts, responses to hypothetical questions, and empathy with certain ideas. Some of these are linked to personality tests while the rest were our invention.

Association with sugar plum fairy		Imagination	
Dancing	23	Go left	7
Food	26	Go right	21
Music	21	Go straight on	54
Others	12	Imagination	
Association with golf		Reverse my car out	15
A car	5	Drive out forwards	23
A sport	76	Not applicable as I don't drive to work	44
Others	1	Empathy – simplicity	
Association with translation		Agree	76
Circle	36	Disagree	6
Square	11	Empathy - translation of songs	
Triangle	15	Agree	63
None	20	Disagree	19
Association with doughnut		Empathy - direction signs	
One with a hole in the middle	74	Yes	50
A spherical one filled with jam etc.	8	No	32
Imagination			
Mars	41		
Venus	41		

Table 11: Personality Features. The left side shows five associational characteristics. For example, the first asks what the main association is which the participant makes with ‘sugar plum fairy’. The right side shows five other similar traits. This results in five multi-valued features, one for each of the associational characteristics and traits, and five Boolean features, three for empathy questions, one for doughnut association, and one for the choice of planet.

4.7 Background

Finally, we elicited background information, for example, language skill, occupation, and qualifications. We also added four attributes during the process of recruiting participants: membership of translator’s associations – in Japan or otherwise, being a student, and being a teacher.

Native language		Other Fluent languages	
Chinese	1	English	46
English	38	French	2
Filipino	2	German	3
French	3	Italian	1
Japanese	34	Japanese	20
Polish	1	Spanish	4
English and French	1	Korean	4
English and Japanese	1	Welsh	1
Korean and Japanese	1	Indonesian	1
		Creole	1

Table 12: Background - Language Skill Features. The first two columns show various languages and how many participants were native speakers of each. The second two columns show the languages in which they were fluent. This results in one multi-valued feature for native language, six Boolean features for fluent languages (English, French, German, Italian, Japanese and Spanish), and one multi-valued feature for the remaining fluent languages.

Languages learned at		Lived in a country using learned language	
Primary school	20	Yes	60
Secondary school	46	No	22
University	62	Language qualification	
Language school	29	English	10
Other	13	Japanese	28
Usage of Japanese		Education	
1 to 5 years	22	Secondary	82
6 to 10 years	5	Bachelor	67
11 to15 years	3	Master	24
16 to 20 years	2	Doctoral	4
More than 20 years	50	Translation study	18
Usage of English		Being a translator	
1 to 5 years	4	1 to 5 years	9
6 to 10 years	10	6 to 10 years	6
11 to15 years	10	11 to15 years	5
16 to 20 years	5	16 to 20 years	3
More than 20 years	53	More than 20 years	9
		Never	50

Table 13: Background - Language Learning Features. The first two columns are concerned with where languages were learned, and how many years experience participants had in Japanese and English respectively. The second two columns deal with whether a participant had lived in the country of a learned language, their qualifications, their education, and their years of experience as a translator. This results in eight Boolean features, one for living in a target language country, one each for language qualification in English and in Japanese, four for educational qualifications, and one for a degree in translation. There are also four multi-valued features, one for the place where a language was learned, one for the use of Japanese, one for the use of English, and one for the years of being a professional translator.

Member of Translators Association	
Yes	28
No	54
Member of Japanese Translators Association	
Yes	12
No	70
Student	
Yes	35
No	47
Teacher	
Yes	11
No	71
Place of Residence	
Japan	33
Ireland	23
USA	12
China	1
France	5
Australia	4
UAE	1
Wales	1
Canada	1
Korea	1
Gender	
Male	32
Female	47
Unknown	3

Table 14: Background - Residence and Occupation Features. This shows how many participants were members of translators associations in English or Japanese, how many were students or teachers, what their place of residence was, and finally what their gender was. This results in four Boolean features and two multi-valued features.

4.8 Summary of Features

We generated 146 attributes under the six categories: Arts, Sports, Pastimes, Lifestyle, Personality, and Background. Concerning Arts (Table 1), many were not familiar with English novels (63/82) or poems (42/82) but the most popular were Bleak House and Shall I compare Thee. Similarly (Table 2), they were mainly not familiar with Japanese novels (51/82) or poems (38/82) but the most popular were those by Natsume and Miyazawa. 64/82 were familiar with at least one work of classical music and 47/82 knew at least one Japanese song. The 5th Symphony of Beethoven (Table 3) was by far the most popular classical work - 34 knew it. 48/82 were familiar with at least one painting (Table 4), The Night Watch being the most popular (24/82).

Turning to Sports (Table 5), 59/82 played no team sport but of those who did, 13 played just one, 5 played two, and 5 played three. The majority (57/82) were familiar with one or more individual sports.

For Pastimes (Table 6), 54/82 liked puzzles, 25 liked one, 19 liked two, 9 liked three, and 1 even liked four. So puzzles are popular. 45/82 liked card games, 16 liked one, 19 liked two, and 6 liked three.

52/82 liked board games, 28 liked one, and 20 liked two. Generally these pastimes were popular with the translators.

46/82 played at least one musical instrument (Table 7) and 29/82 played just one, the most popular being the piano (29/82). 69/82 played in no musical ensemble but 11/82 played in one.

74/82 enjoyed at least one performing art (Table 8), so this was very popular. 15 enjoyed one, 21 two, and 30 three. Rakugo was favoured by 36/82.

Considering newspapers (Table 9), 34/82 read both English and Japanese papers while a further 38 read papers in one language. 16/82 read the Guardian and 24/82 read Asahi.

Concerning Lifestyle (Table 10) only 16/82 grew up in a bilingual environment. While 76/82 had a TV, 56/82 preferred the radio. 67/82 liked to read a book and 70/82 liked bicycling. 78/82 had a friend from abroad and 74/82 used a library. 38/82 had a mechanical pencil sharpener (we were interested in the connection between translation and pencil sharpening). For fruit juices, orange was the most popular (35/82). 74/82 were right-handed. Only 39/82 wore a wristwatch, mostly on the left hand (33/82).

For Personality (Table 11), the sugar plum fairy was fairly equally divided in its association between dancing (23), food (26), and music (21). Golf was however considered primarily a sport by 76/82. Translation had a circular association for 36/82. A doughnut had a hole in the middle for 74/82 - a very high figure. Perhaps this is the main type in the world. For imagination, Mars (41/82) and Venus (41/82) were equal. At a junction, 54/82 went straight on. 44/82 had no car but of the remainder, 23 drove out forwards and 15 reversed out. This was our general expectation - 23/38 i.e. 61% were planning ahead by reversing into their drive the night before. Simplicity was popular (76/82) and inaccurate song translation was disapproved of by most (63/82). 50/82 noticed errors in direction signs.

Concerning language background (Table 12) 38/82 were native English speakers and 34/82 were native Japanese speakers. For language learning (Table 13), most (62/82) learned a language at university and 13 learnt from other sources e.g. Manga. 60/82 had lived in a country using the learned language - a high figure. 50/82 had never worked as a translator.

Finally, concerning residence and occupation (Table 14), most were not members of an English (54/82) or Japanese 70/82) translators association. 35/82 were students and only 11/82 were teachers. The most popular places of residence were Japan (33/82) Ireland (23/82), and USA (12/82) - this probably reflects our recruitment. 47/82 were female (57%).

5. Selection of Texts

We had to find an appropriate text to translate so that we could generate a maximum spread of translation outcomes and also have a reasonable number of translation participants. To set the appropriate number of texts, together with their length and topic, we studied specifications for various translators' accreditation bodies (Table 15). Generally, such bodies use two texts, one general and one specific, and these tend to be quite short - less than 300 words in most cases (ITI is the exception at 1,000 words). For accreditation, a candidate would have a specialisation which would in turn determine the choice of specific text to be used in the second translation. In our case, participants were of many backgrounds, making this difficult to accomplish. In addition, an important factor was to choose texts which were not too easy and which could not therefore be translated online using Google Translate and related tools. To satisfy these criteria we used texts from the proceedings of the House of Representatives of Japan (2014) which includes transcripts of all debates in the Japanese Parliament. Two topics were chosen: English Language Education in Japan (henceforth EE) and Wine Consumption (WC). Sentences were selected which featured idiomatic and rhetorical expressions as well as complex structure. In total the two texts comprised 1,610 moji characters in 25 sentences. Table 16 summarises the characteristics of the sentences; in general, the EE text contained a larger number of short sentences, while WC contained a smaller number of longer sentences.

	No. of texts	Vol. of each text (words)	Duration	Domain
ATA	2	225-275	3 hours	General
CTTIC	2	175-187	3 hours	General/Specific purpose
NAATI	2	140-150	1 hour	General/Specific purpose
ITI	1	1000	5 days	Not specified
ITIA	2	Short/Long	3 hours	Specific purpose
JTF	1	300 words /	2 hours	Specific purpose

Table 15: Evaluation Text Requirements for Translators' Associations

		Character count			
	The number of sentences	Total	The longest sentence	The shortest sentence	Ave.
English Education	16	841	130	12	52.5
Wine Consumption	9	769	192	34	85.4

Table 16: Evaluation Text Specification for this Study

We verified that Google Translate did not produce good translations with the sentences in our dataset. The sentences in our dataset had not been translated before to the best of our knowledge. Therefore, we were confident that the work produced by participants was original and that they were not able to 'cheat'.

6. Translations and their Evaluation

6.1 Translations

Having asked participants to translate the test sentences from Japanese to English, we received 1,822 sentences from 82 participants. A total of 278 sentences were not translated. Some participants made spelling and typing errors. Some errors could have been a simple typing error, but other errors were due to the inexperience of the participant regarding English. For example, Thailand was spelled 'Tailand' or Korea was spelled 'Corea'. Participants were not restricted in the way in which they carried out the translation, except that it was carried out sentence-by-sentence. This resulted in a situation where participants occasionally referred to a concept expressed in the previous sentences. This reference across sentences made the evaluation task difficult. We noticed a typical human characteristic that would be different from machine translation which is illustrated in Table 17. Candidate Translation 3 refers to the criticism that Japanese people are not good at English, something which is expressed in the previous sentence. On the other hand, Candidate Translation 4 refers to it using 'this' alone. Hence, sentences could not be evaluated in isolation.

Example source sentence:「自分自身の英語能力も含めてこう思うんです。」
Candidate translation 3: 'I include my own English ability in this criticism.'
Candidate translation 4: 'This applies to myself as well'
Google Translation: 'And I think this way, including his own English ability.'

Table 17: Human and Machine Translations of a Sample Sentence

We also noticed that some translations contained concepts which were not in the original. How should we score such translations? We return to this in the next section.

6.2 Nugget Recall and Fluency

What an evaluator generally does when analysing a sentence became apparent to us during the previous trials prior to this work. The evaluator examines a sentence from two aspects: 1. Concepts in a source sentence should be present in the translation, 2. Each concept should be described fluently in the translation. This led us to develop an evaluation metric based on the Nugget Recall concept introduced at TREC for the evaluation of definition questions in the Question Answering track (Voorhees 2003). Voorhees determined, for each question, the points which a complete answer should contain; these were termed nuggets. When evaluating a candidate answer, human assessors counted the number of reference nuggets it contained. The Nugget Recall of the answer was then calculated by dividing this value by the total number of reference nuggets. Nugget Precision was more difficult to assess at TREC because it was not clear what information was 'excess' or exactly how to punish it. They chose a method which assumes an allowance of 100 characters for each correct nugget returned by a system. If a candidate answer is no longer than its allowance, it has Nugget Precision 1. If it is longer, it receives a lower precision.

We aimed to adapt the Nugget Recall measure for our task of evaluating translations. For us, a nugget is a representation of a concept that is present in a source sentence and must therefore be present in the translation. An evaluator was informed for each sentence what reference nuggets it contained. They could then judge for each reference nugget whether it was present in the candidate translation or not. Prior to the evaluation, the first author, a native Japanese speaker, determined the reference nuggets for all the source sentences. The nuggets were described in English and the second author, a native English speaker, reviewed each nugget to check that it was understandable and unambiguous.

We engaged four native English speakers who did not speak Japanese to evaluate the quality of the translated sentences. They worked independently and remotely using an online translation evaluation tool developed for the purpose. The sequence of sentences within a topic was randomised and one sentence at a time was presented to the evaluator (Figure 1). The left side of the screen showed the previous sentence in the original text, so that the correctness of anaphors in the target sentence could be judged. Below the sentence, the required nuggets, as determined above, were listed. The evaluator was asked to tick those nuggets which they deemed to be present in the target. After this, they had to judge the fluency on a scale of one to five. Finally, there was a box at the bottom for any comments the evaluator wished to make. Once a fluency level was indicated, 'Next' and 'Previous' buttons appeared. Evaluators were allowed to go back to previous sentences whenever they wished. They could also stop the evaluation process at any time and resume later.

There were 82 participants and 25 sentences, sixteen on the topic of English Education and nine about Wine Consumption. This made a total of 2,050 sentences, each of which was judged four times, once by each evaluator. Following the evaluation process, we had thus obtained, for each sentence, four counts of the number of nuggets judged to be present, and four fluency scores. From this data we computed the average number of nuggets. We then computed the Sentence Nugget Recall as follows:

Sentence Nugget Recall = average no. nuggets found / no. reference nuggets

The Sentence Fluency was defined to be the average of the four fluency scores provided by the evaluators. The harmonic mean of Nugget Recall and Fluency was the translation score of a sentence. The average over the 25 test sentences was defined to be the overall translation score of the participant.

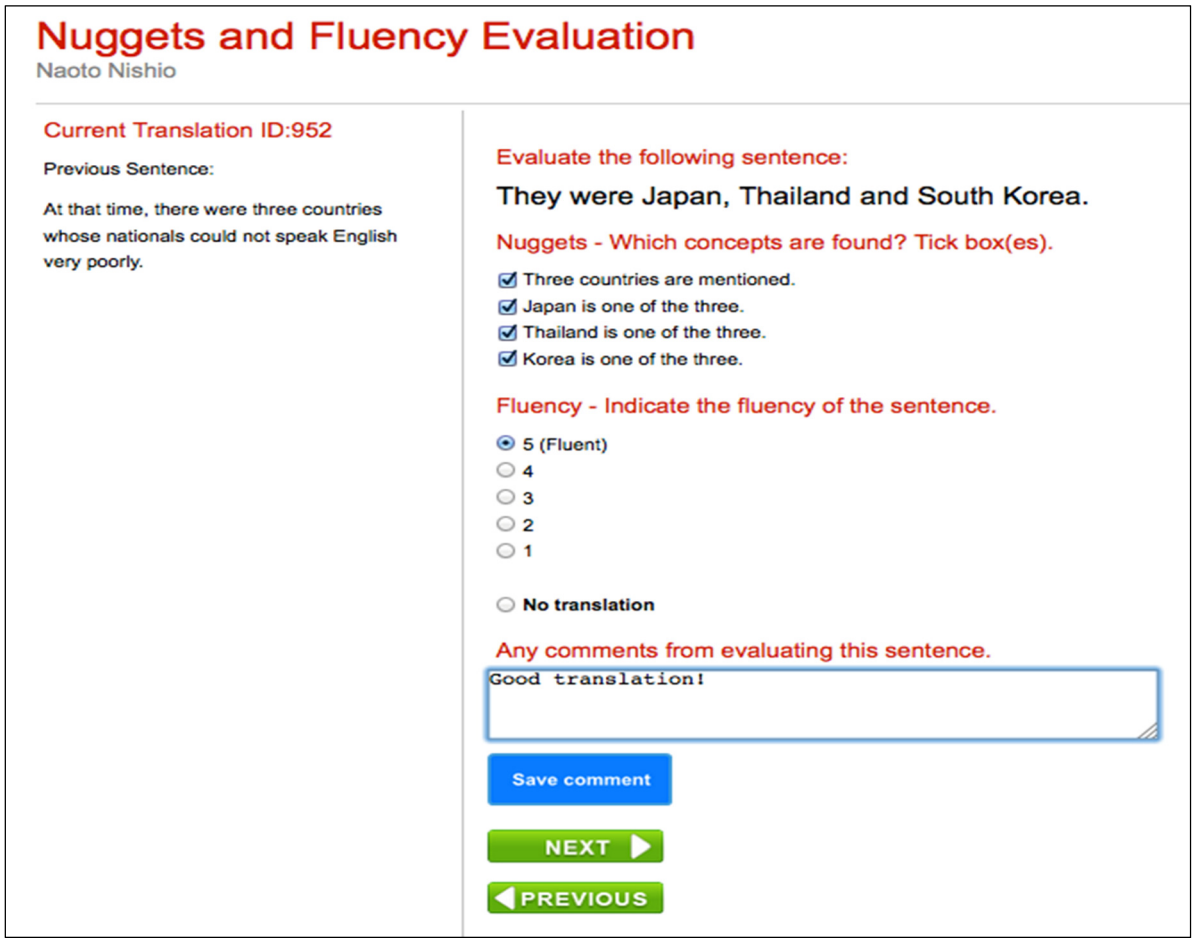

Figure 1: Nugget Recall and Fluency Evaluation Tool

6.3 NRF worked example

We will explain how Nugget Recall works using a sample sentence from the task together with two candidate translations of it produced by participants. Table 18 shows sentence no. 7 from the topic of Wine Consumption (WC). It is the longest sentence by character count of the 25 source sentences and it combines almost seven normal sentences together. It also contains metaphor and simile as well as borrowing a word from English. This is a complex sentence from a political discussion and it was too difficult for Google Translate to translate correctly. Based on the reference translation in Table 18, we defined nine nuggets as shown in Table 19. Four evaluators (E1, E2, E3, and E4) then assessed translations returned by candidates. Table 20 shows examples of two candidate translations. Candidate Translation 1 (CT1) is the best translation in the 82 candidate translations as it scored 0.88 using NRF. Candidate Translation 2 (CT2) is a rather inferior translation that scored 0.69. Table 21 shows the results of Nugget Recall evaluation by the four evaluators. For CT1, three evaluators agreed that the sentence contained all nuggets 72-80 in Table 19, while E4 judged that 78 was missing. The Nugget Recall of the sentence is thus 1 + 1+ 1 + 0.89 / 4 evaluators, i.e. 0.97. Similarly, for CT2, E1 and E2 found all nuggets, E3 found only three nuggets (74,75 and 77), and E4 found only two (77 and 80). Therefore, Nugget Recall of the sentence is 0.64. In cases where a translation is unclear and incomplete, we can expect divergence between evaluators.

Example source sentence:「ただ、米もお酒も消費拡大しろ、ワインも消費拡大しろ、今度は特区構想でどぶろくをつくれ、こうなってまいりますと、日本人の胃袋と肝臓というのは限界がございまして、ワインはかなりそういう意味ではアッパーリミット的なところに定着をしてきたんじゃないか、むしろ世界の中で国産ワインと海外のワインとどう競争していくかという時代になったのではないかなという感じを、今御意見を伺いながらいたしました。」
Reference translation: 'However, being asked to increase the consumption of Japanese sake and wine, as well as being asked to allow unrefined sake in a special district, I must say that there are limits for the stomachs and livers of Japanese nationals, and therefore I came to realise, as I was listening to your opinion, that wine is reaching an upper limit, and we are entering an era where our domestic wine must compete with foreign wine.'
Google Translation: 'However, the US also drink also white consumption expansion, wine white expanding consumption, in turn, make the rice wine in the special zone initiative, and we will come to this, because Japanese of stomach and liver are there is a limit, wine fairly or not we've been fixing the upper limit specific place in that sense, but rather a domestic wine and feeling that kana not became the era of how we compete with foreign wine in the world, and now it has decided while ask for your opinion.'

Table 18: Source Sentence and a Reference Translation with Google Translation

Nugget ID	Nugget description
72	The speaker has been asked to increase the consumption of certain goods.
73	One of which is sake.
74	One of which is rice.
75	One of which is wine.
76	The speaker has been asked to produce Doburoku (unrefined sake).
77	The speaker mentions that he has been listening to an opinion.
78	The speaker feels that wine consumption in Japan has reached its upper limit.
79	The speaker feels that it is time to consider competing in the international wine market by exporting domestic wine.
80	This upper limit is described using a metaphorical expression (stomach and liver).

Table 19: Nine Nugget Definitions for the Example Sentence

Concerning the fluency score of each participant, it was calculated as the average fluency over the 25 sentences. Finally, the overall translation score for a participant is the harmonic mean of their average sentence recall and average sentence fluency, as we have already noted.

Candidate Translation 1: 'However, the reality is that we are being bombarded with requests for rice and sake consumption increases, in addition to wine consumption increase; or an idea of considering the production of doburoku in a special economic zone?, etc., etc. We must remind you that there is a limit the stomach or liver of Japanese people can take, as you know. In that sense, while I was listening to your remarks, I was feeling that wine has possibly reached such an upper limit so that it may be that now is the time to think about how to compete with the products of other nations in the export market.'
Candidate translation 2: 'If the situation occurs that we are told to expand the consumption of rice, alcoholic beverages, as well as wines, and furthermore, to create unrefined sake according to a special zone plan, the consequence could be that Japanese peoples' stomachs and livers could reach their capacity for consumption, and subsequently impose a finite state to wine consumption. Also, as I was listening to your opinions I wondered how domestically produced wines could, in the current era, somehow compete with overseas produced ones.'

Table 20: Two Candidate Translations for the Source Sentence in Table 18

Candidate translation 1				Candidate translation 2			
E1	E2	E3	E4	E1	E2	E3	E4
72	72	72	72	72	72	-	-
73	73	73	73	73	73	-	-
74	74	74	74	74	74	74	-
75	75	75	75	75	75	75	-
76	76	76	76	76	76	-	-
77	77	77	77	77	77	77	77
78	78	78	-	78	78	-	-
79	79	79	79	79	79	-	-
80	80	80	80	80	80	-	80

Table 21: Evaluation Results by Four Evaluators Concerning Two Examples. For example, Evaluator E3 judged that Candidate Translation 1 contained Nugget id 78 while Evaluator E4 judged that it did not.

6.4 Translation Scores of Participants

Table 22 below shows a frequency analysis of the translation scores. Although this analysis was based on a modest set of 82 samples, the results indicated that those who considered themselves translators were clustered rather narrowly at a high quality score region. A total of 35 students were clustered at a low mean value. A high median result from the non-student group showed that there were some who could translate as well as skilled translators. However, the average score for students detracted from the overall result. Based on this analysis, we separated participants into three categories: 'All participants', 'Professional Translators', and 'Non-Professional Translators'. Experiments of the main study were planned using these three.

	All	Translator	Non-translator	Non-student	Student
Valid sample	82	32	50	47	35
Mean	0.60	0.73	0.52	0.71	0.46
Median	0.69	0.77	0.64	0.74	0.55
Mode	0.69	0.83	0.69	0.74	0.25
Std. Deviation	0.22	0.13	0.23	0.12	0.25
Skewness	-1.09	-2.18	-0.66	-1.73	-0.18
Std. Error of Skewness	0.27	0.41	0.34	0.35	0.40
Range	0.82	0.54	0.81	0.54	0.81
Minimum	0.02	0.30	0.02	0.30	0.02
Maximum	0.84	0.84	0.83	0.84	0.83

Table 22: Frequency Analysis of the Translation Scores

We examined the inter-rater reliability of the data using Kripendorff's Alpha (Hayes and Krippendorff 2007; Krippendorff 2011). The resulting Alpha was 0.73 for Nugget Recall and 0.61 for fluency. We considered that these levels were acceptable. The agreement for recall is particularly good, indicating that our NRF measure is robust – once the nuggets for each sentence have been determined.

The lowest average translation score was recorded for Sentence 18: その平成十年を頂点にして頭打ちなんですね、ワインの消費量というのが。This sentence features an inverted structure, idiom expressions, and a reference to the Japanese year; these characteristics may explain why it was so difficult to translate, despite its low word count. Examples of correct translations, together with a literal translation for reference, can be seen in Table 23.

Original Sentence							
その平成十年を頂点にして頭打ちなんですね、ワインの消費量というのが。							
Literal Translation							
その	平成	十年を	頂点にして	頭打ちなんですね、	ワインの	消費量	というのが。
That	Heisei	10 years	as a peak	hit the head	of wine	amount of consumption	which is called

Candidate Translation	Score
Wine consumption reached a peak and then plateaued in 1998.	0.95
Wine consumption peaked and plateaued in 1998 (H.10).	0.85
The wine consumption has reached a plateau since 1998.	0.85
The amount of wine consumption reached the summit in 1998 and after that becomes stagnant.	0.85
During that 10th Year of Heisei, it peaked and plateaued, that is, the consumption volume of wine did.	0.79

Table 23: Difficult Sentence with Good Candidate Translations. The top line shows Sentence 18 from the test set. Below this is a word-for-word translation. At the bottom are five translations made by participants together with their translation scores.

This example suggests that our NRF measure can uncover subtle differences between translations.

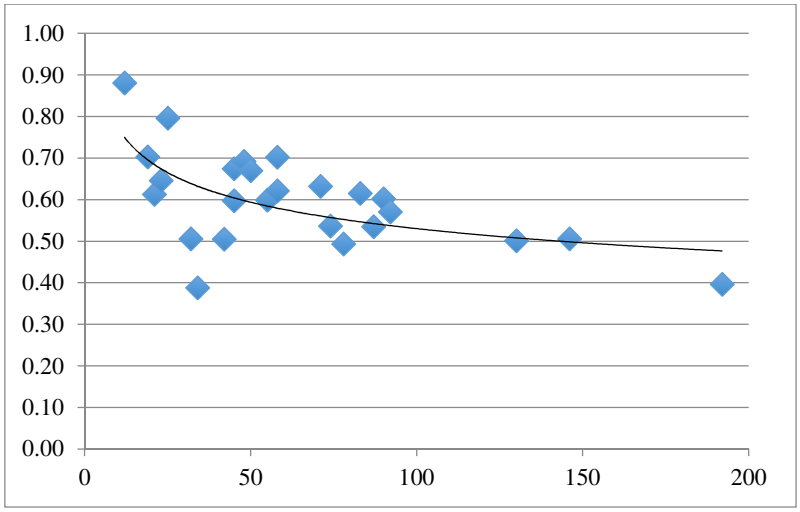

Figure 2: Correlation between Average Translation score and Test Sentence Word Count

Experiment	Number of Participants	#G	#M	#B	Translation Score Threshold(s)	F-weighted Average (classes)	F-Good	F-Bad	Kappa	Attribute Subset
1	82	23	0	59	G ≥ 0.765	0.728	0.500	0.817	0.317	Background
2	82	48	0	34	G ≥ 0.653	0.656	0.748	-	0.288	Background
						0.635	0.724	-	0.244	Arts
						0.619	0.687	-	0.211	Lifestyle
3	82	23	36	23	G ≥ 0.765	0.550	0.511	-	0.317	Lifestyle
					M ≥ 0.554	0.606	0.429	0.727	0.388	Background
4	82	23	0	59	G ≥ 0.765	-	-	-	-	-
		48	0	34	G ≥ 0.653	0.670	0.722	-	0.318	Wrapper
		23	36	23	G ≥ 0.765 M ≥ 0.554	0.658	0.667	-	0.479	Manual
5	47	20	0	27	G ≥ 0.766	-	-	-	-	-
		36	0	11	G ≥ 0.660	0.719	0.649	-	0.423	Wrapper
						0.774	0.687	-	0.541	Manual
6	31	17	0	14	G ≥ 0.767	0.775	0.788	-	0.547	Wrapper
						0.737	0.714	-	0.498	Manual
						0.678	0.688	-	0.356	Lifestyle
						0.710	0.710	-	0.424	Pastime

Table 24: Summary of Six Experiments. The columns #G, #M, and #B are the number of participants for Good, Medium and Bad translation quality, respectively. The column Attribute Subset lists the name of the subset which generated the highest average F-score in the experiment.

7. Experiments

7.1 Outline

A total of six experiments were carried out using the J48 decision tree algorithm on WEKA (Witten 2013) based on the well-known C4.5 algorithm of Quinlan (1993). In all experiments the input was a set of attributes describing the translator, and the output was a classification of that translator. The precise features used and the nature of the classification varied. In particular, we were interested to discover which of the attributes were most important in making the prediction. Table 24 summarises the results. The J48 confidence threshold for pruning was set at 0.25 and the minimum number of instances permissible at a leaf was set at two. All the decision trees were trained and evaluated using 10-fold cross validation (Alpaydin 2004). The factors varying in different experiments were:

- Whether the classification performed by the decision tree was two way or three way; a two-way system assigned a participant to class Good or Bad, depending on whether they were considered to be a good translator or a bad translator. A three-way approach assigned a participant to Good, Medium or Bad. Experiments 1, 2, 5 and 6 were Good/Bad, Experiment 3 was Good/Medium/Bad, and Experiment 4 included both Good/Bad and Good/Medium/Bad.
- What the values of thresholds were; these were used to assign participants to classes (i.e. Good/Bad or Good/Medium/Bad) depending on their translation scores. For example, lowering the translation score threshold for assigning a translator to the Good class would result in a larger number of Good translators. Thresholds were also altered for certain experiments to balance the number of Good and Bad training examples.
- Categories of participants used for training within an experiment; Experiments 1-4 used all participants, Experiment 5 excluded bad translators, and Experiment 6 only included professional translators.
- The means by which features were selected; as discussed earlier, there were six sets of features in total – Arts, Background, Lifestyle, Pastimes, Personality, and Sports – and various subsets of these could be manually chosen to use in a particular experiment. Alternatively, features could be selected automatically using the Wrapper mechanism provided in WEKA.

7.2 Experiment 1

This was a two-way classification: Good and Bad. Translators with a score of 0.8 and over were considered Good (23 in total), and the rest were Bad (59 in total). The system was trained seven times using seven predefined sets of attributes: All attributes, Arts (literature, music, paintings), Sports, Pastimes, Lifestyle, Personality, and Background.

Feature	Classified as		Class	TP rate	FP rate	Precision	Recall	F	ROC Area		Baseline	OneR
	Good	Bad										
All	6	17	Good	0.261	0.203	0.333	0.261	0.293	0.431		0	0.615
	12	47	Bad	0.797	0.739	0.734	0.797	0.764	0.431		0.837	0.880
	k = 0.0616		Weighted Avg.	0.646	0.589	0.622	0.646	0.632	0.431		0.602	0.806
Arts	5	18	Good	0.217	0.237	0.263	0.217	0.238	0.496		0	0.176
	14	45	Bad	0.763	0.783	0.714	0.763	0.738	0.494		0.837	0.785
	k = -0.021		Weighted Avg.	0.610	0.630	0.588	0.610	0.598	0.495		0.602	0.614
Background	11	12	Good	0.478	0.169	0.524	0.478	0.500	0.672		0	0.615
	10	49	Bad	0.831	0.522	0.803	0.831	0.817	0.672		0.837	0.880
	k = 0.3172		Weighted Avg.	0.732	0.423	0.725	0.732	0.728	0.672		0.602	0.806
Lifestyle	5	18	Good	0.217	0.119	0.417	0.217	0.286	0.522		0	0.438
	7	52	Bad	0.881	0.783	0.743	0.881	0.806	0.522		0.837	0.864
	k = 0.1156		Weighted Avg.	0.695	0.596	0.651	0.695	0.660	0.522		0.602	0.744
Pastime	6	17	Good	0.261	0.153	0.400	0.261	0.316	0.545		0	0.812
	9	50	Bad	0.847	0.739	0.746	0.847	0.794	0.546		0.837	0
	k = 0.1212		Weighted Avg.	0.683	0.575	0.649	0.683	0.660	0.546		0.602	0.584
Personality	3	20	Good	0.130	0.051	0.500	0.130	0.207	0.505		0	0.138
	3	56	Bad	0.949	0.870	0.737	0.949	0.830	0.505		0.837	0.815
	k = 0.1028		Weighted Avg.	0.720	0.640	0.670	0.720	0.655	0.505		0.602	0.635
Sports	0	23	Good	0	0	0	0	0	0.446		0	0
	0	59	Bad	1	1	0.720	1	0.837	0.446		0.837	0.829
	k = 0		Weighted Avg.	0.720	0.720	0.518	0.720	0.602	0.446		0.602	0.596

Table 25: Results of Experiment 1

Results are in Table 25. Statistical significance tests for all experiments can be found in Section 8.1, Table 32. The classification ability, determined by F-Measure on Good, was generally weak. The best result (0.5) was using the Background feature set. However, the Bad class returned 0.817 and F-Measure of the weighted average between the two classes was 0.728. As the training data with the chosen threshold of 0.8 comprised 23 Good and 59 Bad examples, it was thus skewed to the latter. This could have affected the result.

7.3 Experiment 2

This was a two-way classification: Good and Bad. This time, translators with a score of 0.7 and over were considered Good (48 in total), and the rest were Bad (34 in total). Once again, the system was trained seven times using seven predefined sets of attributes: All attributes, Arts (literature, music, paintings), Sports, Pastimes, Lifestyle, Personality, and Background.

Feature	Classified as		Class	TP rate	FP rate	Precision	Recall	F	ROC Area		Baseline	OneR
	Good	Bad										
All	22	26	Good	0.458	0.500	0.564	0.458	0.506	0.520		0.738	0.727
	17	17	Bad	0.500	0.542	0.395	0.500	0.442	0.520		0	0.233
	k = -0.0401		Weighted Avg.	0.476	0.517	0.494	0.476	0.479	0.520		0.432	0.522
Arts	38	10	Good	0.792	0.706	0.613	0.792	0.691	0.471		0.738	0.500
	19	10	Bad	0.294	0.208	0.600	0.441	0.508	0.631		0	0.421
	k = 0.2422		Weighted Avg.	0.585	0.500	0.566	0.585	0.558	0.470		0.432	0.467
Background	40	8	Good	0.833	0.559	0.678	0.833	0.748	0.627		0.738	0.735
	19	15	Bad	0.441	0.167	0.652	0.441	0.526	0.627		0	0.606
	k = 0.2881		Weighted Avg.	0.671	0.396	0.667	0.671	0.656	0.627		0.432	0.681
Lifestyle	34	14	Good	0.708	0.500	0.667	0.708	0.687	0.628		0.738	0.742
	17	17	Bad	0.500	0.292	0.548	0.500	0.523	0.628		0	0.627
	k = 0.211		Weighted Avg.	0.622	0.414	0.618	0.622	0.619	0.628		0.432	0.694
Pastime	25	23	Good	0.521	0.588	0.556	0.521	0.538	0.481		0.738	0.619
	20	14	Bad	0.412	0.479	0.378	0.412	0.394	0.480		0	0.157
	k = -0.0665		Weighted Avg.	0.476	0.543	0.482	0.476	0.478	0.481		0.432	0.428
Personality	40	8	Good	0.833	0.882	0.571	0.833	0.678	0.500		0.738	0.655
	30	4	Bad	0.118	0.167	0.333	0.118	0.174	0.498		0	0.167
	k = -0.0541		Weighted Avg.	0.537	0.586	0.473	0.537	0.469	0.499		0.432	0.453
Sports	37	11	Good	0.771	0.912	0.544	0.771	0.638	0.354		0.738	0.727
	31	3	Bad	0.088	0.229	0.214	0.088	0.125	0.354		0	0.233
	k = -0.1542		Weighted Avg.	0.488	0.629	0.407	0.488	0.425	0.354		0.432	0.522

Table 26: Results of Experiment 2

Results are in Table 26. In general, the classification ability of the Good class was improved from that of Experiment 1. The best result was once again from the Background feature set: The F-score of the Good class was 0.748 and the weighted average F-Measure was 0.656. In addition, the feature sets Arts and Lifestyle returned reasonably good F-Measure for the Good class: 0.691 and 0.687 respectively. The weighted average F-Measures were 0.558 and 0.619 respectively. So this supports our hypothesis that Arts and Lifestyle characteristics can influence the quality of a translator.

```
lit_poem_miyazawa_ja = FALSE
|   lit_poem_familiar_preference_ja = Not familiar with any of these
|   |   music_bach_eu = FALSE
|   |   |   lit_familiar_preference_en = Brideshead Revisited - Waugh: bad (2.02)
|   |   |   lit_familiar_preference_en = Not familiar with any of these
|   |   |   |   lit_poem_familiar_preference_en = Shall … a summers day - Shakespeare: bad (4.05)
|   |   |   |   lit_poem_familiar_preference_en = Tyger… Burning Bright - Blake: good (1.01/0.01)
|   |   |   |   lit_poem_familiar_preference_en = Not familiar with any of these: good (9.11/1.11)
|   |   |   |   lit_poem_familiar_preference_en = You … Father William - Carroll: good (2.02/0.02)
|   |   |   lit_familiar_preference_en = Bleak House - Dickens: good (2.02/0.02)
|   |   |   lit_familiar_preference_en = To the Lighthouse - Woolf: good (1.01/0.01)
|   |   music_bach_eu = TRUE
|   |   |   music_preference_eu = Beethoven - 5th Symphony
|   |   |   |   music_tonbo_ja = FALSE: bad (4.05)
|   |   |   |   music_tonbo_ja = TRUE: good (2.02/0.02)
|   |   |   music_preference_eu = Tchaikovsky - Piano Concerto No.1: bad (7.09)
|   |   |   music_preference_eu = Not familiar with any of these: bad (0.0)
|   |   |   music_preference_eu = Takemitsu - A Flock Descends … Pentagonal Garden: bad (0.0)
|   |   |   music_preference_eu = Bach - Magnificat: good (1.01/0.01)
|   |   |   music_preference_eu = Faure - Elegie for Cello: bad (0.0)
|   lit_poem_familiar_preference_ja = Hagiwara: bad (4.05)
|   lit_poem_familiar_preference_ja = Miyazawa: good (1.01/0.01)
|   lit_poem_familiar_preference_ja = Yosano: bad (2.02)
lit_poem_miyazawa_ja = TRUE: good (39.48/9.48)
```

Figure 3: Decision Tree for Arts in Experiment 2

```
background_b_uni = FALSE
|   backgrond_lang_qual_en = FALSE: bad (13.0)
|   backgrond_lang_qual_en = TRUE: good (2.0)
background_b_uni = TRUE
|   background_flulang_fr = FALSE: good (64.97/18.97)
|   background_flulang_fr = TRUE: bad (2.03)
```

Figure 4: Decision Tree for Background in Experiment 2

```
music_ensembles_rockpop = FALSE
|   newspaper_read_nikkei = FALSE
|   |   music_instrument_violin = FALSE
|   |   |   pastime_view_rakugo = No
|   |   |   |   pastime_view_film = TRUE: bad (26.32/5.0)
|   |   |   |   pastime_view_film = FALSE: good (4.05/0.05)
|   |   |   pastime_view_rakugo = Yes
|   |   |   |   newspaper_ja_familar = Not familiar with any of these: bad (7.09/1.0)
|   |   |   |   newspaper_ja_familar = Yomiuri: good (3.04/0.04)
|   |   |   |   newspaper_ja_familar = Asahi: good (8.1/1.1)
|   |   |   |   newspaper_ja_familar = Nikkei: good (0.0)
|   |   |   |   newspaper_ja_familar = Mainichi: bad (2.02/1.0)
|   |   |   |   newspaper_ja_familar = Sankei: good (1.01/0.01)
|   |   |   |   newspaper_ja_familar = Tokyo sports: good (1.01/0.01)
|   |   |   |   newspaper_ja_familar = Chunichi: good (1.01/0.01)
|   |   music_instrument_violin = TRUE: bad (2.02)
|   newspaper_read_nikkei = TRUE: good (20.25/2.25)
music_ensembles_rockpop = TRUE: good (6.07/0.07)
```

Figure 5: Decision Tree for Lifestyle in Experiment 2

Looking at the decision trees of Background, Arts, and Lifestyle, we can observe what attributes contributed to the prediction of Good participants (see Figures 3, 4 and 5). Figure 3 shows that familiarity with the Bach Magnificat and the Japanese poem「雨にも負けず」(Miyazawa) are influential; 30 out of the 48 Good participants were familiar with the poem. Fourteen of the Good class were not familiar with the Magnificat or with English novels, but they indicated a particular preference for English poetry. The Background decision tree (Figure 4) indicates that a large number of Good participants had a Bachelor's degree. A total of 13 were predicted as Bad if they did not have either a degree or an English language qualification. The Lifestyle decision tree (Figure 5) shows playing in a rock band, reading the Nikkei newspaper and playing Rakugo as being important.

7.4 Experiment 3

This was a three-way classification: Good/Medium/Bad. Translators with a score of 0.8 and over were considered Good (23 in total), those with a score of 0.554 or less were Bad (23 in total), and the rest were Medium (36 in total). The numbers of Good and Bad were thus balanced, which they were not in Experiments 1 and 2. Once again, the system was trained seven times using seven predefined sets of attributes: All attributes, Arts (literature, music, paintings), Sports, Pastimes, Lifestyle, Personality, and Background.

Feature	Classified as			Class	TP rate	FP rate	Precision	Recall	F	ROC Area		Baseline	OneR
	Good	Med	Bad										
All	9	11	3	Good	0.391	0.305	0.333	0.391	0.360	0.517		0	0.615
	10	20	6	Medium	0.556	0.413	0.513	0.556	0.533	0.549		0.610	0.667
	8	8	7	Bad	0.304	0.153	0.438	0.304	0.359	0.601		0	0.679
	k = 0.1291			Weighted Avg.	0.439	0.310	0.441	0.439	0.436	0.555		0.268	0.656
Arts	5	10	8	Good	0.217	0.153	0.357	0.217	0.270	0.521		0	0.324
	6	26	4	Medium	0.722	0.435	0.565	0.722	0.634	0.615		0.610	0.479
	3	10	10	Bad	0.435	0.203	0.455	0.435	0.444	0.587		0	0.464
	k = 0.2071			Weighted Avg.	0.500	0.291	0.476	0.500	0.479	0.581		0.268	0.431
Background	9	12	2	Good	0.391	0.169	0.474	0.391	0.429	0.689		0	0.615
	8	25	3	Medium	0.694	0.370	0.595	0.694	0.641	0.682		0.610	0.667
	2	5	16	Bad	0.696	0.085	0.762	0.696	0.727	0.716		0	0.679
	k = 0.3886			Weighted Avg.	0.610	0.234	0.608	0.610	0.606	0.694		0.268	0.656
Lifestyle	12	5	6	Good	0.522	0.203	0.500	0.522	0.511	0.692		0	0.133
	9	19	8	Medium	0.528	0.239	0.633	0.528	0.576	0.664		0.610	0.585
	3	6	14	Bad	0.609	0.237	0.500	0.609	0.549	0.664		0	0.692
	k = 0.3179			Weighted Avg.	0.549	0.229	0.559	0.549	0.550	0.672		0.268	0.489
Pastime	4	11	8	Good	0.174	0.254	0.211	0.174	0.190	0.425		0	0.067
	9	23	4	Medium	0.639	0.478	0.511	0.639	0.568	0.513		0.610	0.569
	6	11	6	Bad	0.261	0.203	0.333	0.261	0.293	0.522		0	0
	k = 0.0553			Weighted Avg.	0.402	0.338	0.377	0.402	0.385	0.491		0.268	0.268
Personality	3	15	5	Good	0.130	0.271	0.158	0.130	0.143	0.443		0	0.063
	11	20	5	Medium	0.556	0.609	0.417	0.556	0.476	0.421		0.610	0.515
	5	13	5	Bad	0.217	0.169	0.333	0.217	0.263	0.485		0	0.114
	k = -0.0508			Weighted Avg.	0.341	0.391	0.321	0.341	0.323	0.445		0.268	0.276
Sports	2	14	7	Good	0.087	0.136	0.200	0.087	0.121	0.502		0	0
	4	26	6	Medium	0.722	0.522	0.520	0.722	0.605	0.595		0.610	0.587
	4	10	9	Bad	0.391	0.220	0.409	0.391	0.400	0.508		0	0.250
	k = 0.1189			Weighted Avg.	0.451	0.329	0.399	0.451	0.412	0.545		0.268	0.328

Table 27: Results of Experiment 3

Results are in Table 27. The classification ability of Good was not improved from that of Experiments 1 or 2. The Medium class had the best results across training, perhaps because it had the largest number of training examples.

7.5 Experiment 4

This involved the classifications of Experiments 1-3, namely two-way classification (Good/Bad) and a threshold of 0.8 (from Experiment 1), two-way classsification and a threshold of 0.7 (from Experiment 2), and a three-way classification using 0.8 and over for Good, 0.554 or less as Bad and the rest Medium (Experiment 3). The aim here was to see whether results improved if features were selected automatically or manually. The method used was Wrapper feature selection, i.e.

ClassifierSubsetEval with linear forward selection on WEKA, and Manual selection. For Manual feature selection, we ran J48 on Weka for each feature set, recorded the feature in the top node and amalgamated these to form the feature set.

Feature	Classified as		Class	TP Rate	FP rate	Precision	Recall	F	ROC Area		Baseline	OneR
	Good	Bad										
Manual selection Skewed to Bad	11	12	Good	0.478	0.169	0.524	0.478	0.500	0.672		0	0.500
	10	49	Bad	0.831	0.522	0.803	0.831	0.817	0.672		0.837	0.817
	k = 0.3172		Weighted Avg.	0.732	0.423	0.725	0.732	0.728	0.672		0.602	0.728
Wrapper selection Skewed to Bad	9	14	Good	0.391	0.102	0.600	0.391	0.474	0.620		0	0.378
	6	53	Bad	0.898	0.609	0.791	0.898	0.841	0.620		0.837	0.819
	k = 0.324		Weighted Avg.	0.756	0.466	0.737	0.756	0.738	0.620		0.602	0.695
Wrapper selection Background Skewed to Bad	11	12	Good	0.478	0.153	0.550	0.478	0.512	0.672		0	0.615
	9	50	Bad	0.847	0.522	0.806	0.847	0.826	0.672		0.837	0.880
	k = 0.3392		Weighted Avg.	0.744	0.418	0.735	0.744	0.738	0.672		0.602	0.806
Manual selection Skewed to Good	46	2	Good	0.958	0.676	0.667	0.958	0.786	0.599		0.738	0.761
	23	11	Bad	0.324	0.042	0.846	0.324	0.468	0.599		0	0.471
	k = 0.3098		Weighted Avg.	0.695	0.413	0.741	0.695	0.654	0.599		0.432	0.641
Wrapper selection Skewed to Good	38	10	Good	0.792	0.412	0.731	0.792	0.760	0.737		0.738	0.763
	14	20	Bad	0.588	0.208	0.667	0.588	0.625	0.737		0	0.657
	k = 0.3188		Weighted Avg.	0.707	0.327	0.704	0.707	0.704	0.737		0.432	0.719
Wrapper selection Background Skewed to Good	46	2	Good	0.958	0.676	0.667	0.958	0.786	0.600		0.738	0.763
	23	11	Bad	0.324	0.042	0.846	0.324	0.468	0.600		0	0.657
	k = 0.3098		Weighted Avg.	0.695	0.413	0.741	0.695	0.654	0.600		0.432	0.719
	Good	Med.	Bad									
Manual selection Balanced class	17	1	5	Good	0.739	0.186	0.607	0.739	0.667	0.775		0	0.500
	9	24	3	Med.	0.667	0.196	0.727	0.667	0.696	0.719		0.610	0.584
	2	8	13	Bad	0.565	0.136	0.619	0.565	0.591	0.706		0	0.129
	k = 0.4792		Weighted Avg.		0.659	0.176	0.663	0.659	0.658	0.731		0.268	0.433
Wrapper selection Balanced class	12	8	3	Good	0.522	0.153	0.571	0.522	0.545	0.692		0	0.727
	1	28	7	Med.	0.778	0.304	0.667	0.778	0.718	0.774		0.610	0.711
	2	6	15	Bad	0.652	0.068	0.789	0.652	0.714	0.748		0	0.794
	k = 0.4842		Weighted Avg.		0.671	0.195	0.674	0.671	0.669	0.744		0.268	0.656
Wrapper selection Background Balanced class	12	9	2	Good	0.522	0.085	0.706	0.522	0.600	0.713		0	0.615
	4	27	5	Med.	0.750	0.391		0.750	0.667	0.685		0.610	0.667
	1	9	13	Bad	0.565	0.119	0.650	0.565	0.605	0.690		0	0.679
	k = 0.4216		Weighted Avg.		0.634	0.229	0.644	0.634	0.631	0.694		0.268	0.656

Table 28: Results of Experiment 4

Results are in Table 28. Remember that 'skewed to bad' in the first column of the table refers to the threshold of Experiment 1 and 'skewed to good' refers to that of Experiment 2; these are two-way classifications. 'Balanced' refers to the thresholds of Experiment 3: these are three-way classifications. The best weighted F-Measure (0.738) was from 'Wrapper selection Skewed to Bad' and 'Wrapper selection Background Skewed to Bad'. However, both have low Precision on Good (0.600 and 0.550). The best Precision on Good (0.731) was from Wrapper selection applied on the

dataset from Experiment 2 (Skewed To Good). F-Measure of Good was 0.760 and weighted F-Measure was 0.704. 'Manual selection Skewed to Good' and 'Wrapper selection Background Skewed to Good' also showed high F-Measures at 0.786 for the Good class. However, the classification results of these two datasets were only based on one attribute. Of the three-way classifications, 'Manual selection Balanced Class' was best at Good (F-Measure 0.667).

```
background_member_asso_trans = FALSE
|   painting_rembrandt = FALSE: mid (32.6/9.0)
|   painting_rembrandt = TRUE
|   |   newspaper_read_dailyteleg = FALSE: bad (19.36/6.36)
|   |   newspaper_read_dailyteleg = TRUE: mid (2.04/1.0)
background_member_asso_trans = TRUE: good (28.0/11.0)
```

Figure 6: Decision Tree for Manual Selection Balanced Class in Experiment 4

```
newspaper_read_nikkei = FALSE
|   background_student = TRUE
|   |   background_plc_residence = Ireland: bad (16.0/1.0)
|   |   background_plc_residence = France: bad (4.0/1.0)
|   |   background_plc_residence = Japan
|   |   |   background_wristwatch = Yes on the left wrist: bad (5.0)
|   |   |   background_wristwatch = No wristwatch: good (5.0)
|   |   |   background_wristwatch = Yes on the right wrist: good (1.0)
|   |   background_plc_residence = Rep. of Korea: bad (0.0)
|   |   background_plc_residence = USA: bad (0.0)
|   |   background_plc_residence = UAE: bad (0.0)
|   |   background_plc_residence = Wales: bad (0.0)
|   |   background_plc_residence = Canada: bad (0.0)
|   |   background_plc_residence = Australia: bad (0.0)
|   |   background_plc_residence = China: good (1.0)
|   background_student = FALSE
|   |   choice_crossroad = Go straight on
|   |   |   background_live_trgt_cntry = No: bad (3.08/1.0)
|   |   |   background_live_trgt_cntry = Yes: good (14.39/2.39)
|   |   choice_crossroad = Go right: good (6.17/1.17)
|   |   choice_crossroad = Go left: bad (3.08)
newspaper_read_nikkei = TRUE: good (23.28/2.28)
```

Figure 7: Decision Tree for Wrapper Selection Skewed to Good in Experiment 4

Figures 6 and 7 are two decision trees: one for 'Wrapper selection Skewed to Good' and the other for 'Manual selection Balanced class'. Figure 6 shows that membership of a translator's association predicted 17 Good participants but 11 of the Bad or Medium participants were misclassified. Furthermore, familiarity with paintings and with the Daily Telegraph newspaper separated Medium participants from Bad ones.

Figure 7 shows that familiarity with the Nikkei (a Japanese newspaper), the status of being a student, and the experience of living in a target language country were influential attributes for the prediction of the Good class. A total of 18 Good translators were not students. Students who do not wear a wristwatch and who live in Japan or China produced good translations. Those who read a Japanese Newspaper and those who live in Japan performed well, as did those who live in a country where their

learned language is spoken. This may be an indication of the correlation between reading comprehension of the source text and the quality of translation.

Interestingly, the psychometric questions, for example, choice of a direction at a cross road, were also influential: Good translators tended to go straight or right, and not left. OneR showed that 'Place of residence' and 'Membership of a translators association' were influential.

7.6 Experiment 5

This was a two-way classification: Good/Bad. Seventeen student participants who were exceptionally bad at the translation task were excluded from the training. The number of participants was thus reduced from 82 to 47. Two thresholds were used, 0.8 (from Experiment 1) and 0.7 (from Experiment 2). The 0.8 threshold resulted in 20 Good and 27 Bad training examples, while the 0.7 threshold resulted in 36 Good and 11 Bad examples. For each threshold, seven classifications were carried out with the usual feature sets: All attributes, Arts (literature, music, paintings), Sports, Pastimes, Lifestyle, Personality, and Background. In addition Wrapper feature selection was used. The results are in Table 29 (threshold 0.8, Skewed to Bad) and Table 30 (threshold 0.7, Skewed to Good).

Feature	Classified as		Class	TP rate	FP Rate	Precision	Recall	F	ROC Area		Baseline	OneR
	Good	Bad										
All	11	9	Good	0.550	0.481	0.458	0.550	0.500	0.513		0	0.541
	13	14	Bad	0.519	0.450	0.609	0.519	0.560	0.517		0.730	0.702
	k = 0.0668		Weighted Avg.	0.532	0.463	0.545	0.532	0.534	0.515		0.419	0.633
Arts	4	16	Good	0.200	0.407	0.267	0.200	0.229	0.338		0	0.391
	11	16	Bad	0.593	0.800	0.500	0.593	0.542	0.338		0.730	0.417
	k = -0.2144		Weighted Avg.	0.426	0.633	0.401	0.426	0.409	0.338		0.419	0.406
Background	9	11	Good	0.450	0.296	0.529	0.450	0.486	0.585		0	0.611
	8	19	Bad	0.704	0.550	0.633	0.704	0.667	0.585		0.730	0.759
	k = 0.1568		Weighted Avg.	0.596	0.442	0.589	0.596	0.590	0.585		0.419	0.696
Lifestyle	7	13	Good	0.350	0.222	0.538	0.350	0.424	0.506		0	0.588
	6	21	Bad	0.778	0.650	0.618	0.778	0.689	0.511		0.730	0.767
	k = 0.1339		Weighted Avg.	0.596	0.468	0.584	0.596	0.576	0.509		0.419	0.691
Pastime	7	13	Good	0.350	0.111	0.700	0.350	0.467	0.564		0	0.375
	3	24	Bad	0.889	0.650	0.649	0.889	0.750	0.564		0.730	0.677
	k = 0.2554		Weighted Avg.	0.660	0.421	0.671	0.660	0.629	0.564		0.419	0.549
Personality	5	15	Good	0.250	0.037	0.833	0.250	0.385	0.427		0	0.143
	1	26	Bad	0.963	0.750	0.634	0.963	0.765	0.426		0.730	0.636
	k = 0.2342		Weighted Avg.	0.660	0.447	0.719	0.660	0.603	0.426		0.419	0.426
Sports	5	15	Good	0.250	0.407	0.313	0.250	0.278	0.395		0	0.343
	11	16	Bad	0.593	0.750	0.516	0.593	0.552	0.395		0.730	0.610
	k = -0.2308		Weighted Avg.	0.447	0.604	0.429	0.447	0.435	0.395		0.419	0.496
Wrapper	12	8	Good	0.600	0.185	0.706	0.600	0.649	0.683		0	0.667
	5	22	Bad	0.815	0.400	0.733	0.815	0.772	0.683		0.730	0.793
	k = 0.423		Weighted Avg.	0.723	0.309	0.722	0.723	0.719	0.683		0.419	0.739
Manual	11	9	Good	0.550	0.037	0.917	0.550	0.687	0.680		0	0.421
	1	26	Bad	0.963	0.450	0.743	0.963	0.839	0.680		0.730	0.607
	k = 0.541		Weighted Avg.	0.787	0.274	0.817	0.787	0.774	0.680		0.419	0.528

Table 29: Results of Experiment 5, Skew to Bad

Feature	Classified as		Class	TP Rate	FP Rate	Precision	Recall	F	ROC Area		Baseline	OneR
	Good	Bad										
All	30	6	Good	0.833	0.818	0.769	0.833	0.800	0.563		0.867	0.779
	9	2	Bad	0.182	0.167	0.250	0.182	0.211	0.566		0	0
	k = 0.0167		Weighted Avg.	0.681	0.666	0.648	0.681	0.662	0.564		0.664	0.597
Arts	34	2	Good	0.944	0.909	0.773	0.944	0.850	0.473		0.867	0.821
	10	1	Bad	0.091	0.056	0.333	0.091	0.143	0.461		0	0.125
	k = 0.0473		Weighted Avg.	0.745	0.709	0.670	0.745	0.684	0.471		0.664	0.658
Background	36	0	Good	1	1	0.766	1	0.867	0.549		0.867	0.864
	11	0	Bad	0	0	0	0	0	0.549		0	0.154
	k = 0		Weighted Avg.	0.766	0.766	0.587	0.766	0.664	0.549		0.664	0.698
Lifestyle	32	4	Good	0.889	0.909	0.762	0.889	0.821	0.481		0.867	0.846
	10	1	Bad	0.091	0.111	0.200	0.091	0.125	0.486		0	0.250
	k = -0.0249		Weighted Avg.	0.702	0.722	0.630	0.702	0.658	0.482		0.664	0.707
Pastime	35	1	Good	0.972	1	0.761	0.972	0.854	0.408		0.867	0.867
	11	0	Bad	0	0.028	0	0	0	0.408		0	0
	k = -0.0406		Weighted Avg.	0.745	0.772	0.583	0.745	0.654	0.408		0.664	0.664
Personality	36	0	Good	1	1	0.766	1	0.867	0.428		0.867	0.861
	11	0	Bad	0	0	0	0	0	0.428		0	0.267
	k = 0		Weighted Avg.	0.766	0.766	0.587	0.766	0.664	0.428		0.664	0.722
Sports	34	2	Good	0.944	0.909	0.773	0.944	0.850	0.476		0.867	0.883
	10	1	Bad	0.091	0.056	0.333	0.091	0.143	0.476		0	0.471
	k = 0.0473		Weighted Avg.	0.745	0.709	0.670	0.745	0.684	0.476		0.664	0.787
Wrapper	34	2	Good	0.944	1	0.756	0.944	0.840	0.490		0.867	0.878
	11	0	Bad	0	0.056	0	0	0	0.490		0	0.167
	k = -0.0776		Weighted Avg.	0.723	0.779	0.579	0.723	0.643	0.490		0.664	0.712
Manual	34	2	Good	0.944	0.909	0.773	0.944	0.850	0.548		0.867	0.864
	10	1	Bad	0.091	0.056	0.333	0.091	0.143	0.548		0	0.514
	k = 0.0473		Weighted Avg.	0.745	0.709	0.670	0.745	0.684	0.548		0.664	0.698

Table 30: Results of Experiment 5, Skew to Good

The Manual and Wrapper-selected attribute sets in Skew to Bad returned respectable results (Table 29). The weighted average F-Measures for all of the feature sets exceeded the baseline. The Wrapper-selected feature set returned an F-Measure for Good of 0.649, and a weighted average between two classes of 0.719. Manual feature selection returned a better result: F-Measure for Good of 0.687, and a weighted average between two classes of 0.774.

The results for Skew to Good were disappointing (Table 30). Arts, Sports, and Wrapper-selected feature sets marginally exceeded the baseline. Several instances of the Bad class were misclassified into Good for most of the trainings.

J48:

personality_view_simple_good = Agree
| personality_choice_planet = Venus
| | personality_choice_driveway = Drive out forwards: bad (6.13/2.0)
| | personality_choice_driveway = Not applicable as I do not drive to work
| | | personality_choice_crossroad = Go straight on: bad (8.17/3.0)
| | | personality_choice_crossroad = Go right: good (2.04/0.04)
| | | personality_choice_crossroad = Go left: bad (2.04)
| | personality_choice_driveway = Reverse my car out: good (5.11/0.11)
| personality_choice_planet = Mars: bad (18.39/4.0)
personality_view_simple_good = Disagree: good (5.11/1.11)

OneR:

personality_view_sugerplumfairy:
Other → good
Dancing → bad
Music → bad
Food → bad

Figure 8: J48 Decision Tree and OneR Selected Feature for Personality Features Skewed to Bad in Experiment 5

background_use_of_ja_length = 16 to 20 years: bad (2.04/1.0)
background_use_of_ja_length = lifetime
| background_live_trgt_cntry = Yes
| | view_simple_good = Agree
| | | background_plc_residence = Wales: bad (0.0)
| | | background_plc_residence = Japan: bad (10.0/1.0)
| | | background_plc_residence = USA: good (5.41/1.41)
| | | background_plc_residence = France: bad (0.0)
| | | background_plc_residence = Ireland: bad (1.0)
| | | background_plc_residence = UAE: bad (0.0)
| | | background_plc_residence = Australia: good (3.0)
| | | background_plc_residence = Canada: bad (0.0)
| | view_simple_good = Disagree: good (3.07/0.07)
| background_live_trgt_cntry = No: bad (6.13)
background_use_of_ja_length = 11 to15 years: bad (3.07)
background_use_of_ja_length = 1 to 5 years: bad (2.04)
background_use_of_ja_length = 6 to 10 years: bad (2.04)
background_use_of_ja_length = More than 20 years: good (9.2/1.2)

OneR:
background_plc_residence:
Wales → bad
Japan → bad
USA → good
France → bad
Ireland → bad
UAE → bad
Australia → good
Canada → good

Figure 9: J48 Decision Tree and OneR Selected Feature for Wrapper Selection Skewed to Bad in Experiment 5

J48:
newspaper_read_dailyteleg = FALSE
| pastime_card_bridge = FALSE: bad (33.72/8.0)
| pastime_card_bridge = TRUE: good (6.13/1.13)
newspaper_read_dailyteleg = TRUE: good (7.15/0.15)

OneR:
newspaper_read_dailyteleg:
FALSE → bad
TRUE → good

Figure 10: J48 Decision Tree and OneR Selected Feature for Manual Selection Skewed to Bad in Experiment 5

Returning to Skew to Bad, we will now examine the results produced by the J48 and OneR algorithms. Based on Accuracy, for J48, the highest value is Manual at 0.787 followed by Wrapper at 0.723, and Personality at 0.660 (see Table 32, Section 8.1 for Accuracy scores). The performance of OneR on these is 0.532 on Manual, 0.745 on Wrapper, and 0.489 on Personality. So, OneR is worse

on Manual, better on Wrapper, and worse on Personality. The best accuracy score overall is still J48 on Manual. Wrapper is the only case where both algorithms score highly (J48 Wrapper = 0.723, OneR Wrapper = 0.745). From the point of view of the features being selected, J48 Manual, J48 Wrapper, and OneR Wrapper are the most important to consider, followed by J48 Personality. The other results are low. We now consider the strong results in turn:

The decision tree for J48 Manual (Figure 10) shows newspaper_read_dailyteleg followed by pastime_card_bridge as the most important. This is the highest performing decision tree in Experiment 5 (but not overall in the study). OneR was low here. So, reading newspapers and pastimes have been found to be very important in this investigation. The accuracy level here was statistically significant ($p < 0.05$).

The decision tree for J48 Wrapper (Figure 9) shows background_use_of_ja_length followed by background_live_trgt_cntry, followed by background_plc_residence. Meanwhile, OneR also selects background_plc_residence. So, place of residence is very important. These accuracy levels are significant ($p < 0.05$).

Finally, the decision tree for J48 Personality (Figure 8) shows personality_view_simple_good, personality_choice_planet, personality_choice_driveway, personality_choice_crossroad. The accuracy level here is significant ($p < 0.05$). OneR was low here. This result suggests that these personality traits are worthy of further investigation.

7.7 Experiment 6

This was a two-way classification: Good/Bad. Only professional translators were used with a Good threshold of 0.8. This resulted in 17 Good and 14 Bad translators. In addition to the usual seven attribute subsets, Manual feature selection and Wrapper feature selection were also tried.

Feature	Classified as		Class	TP rate	FP rate	Precision	Recall	F	ROC Area		Baseline	OneR
	Good	Bad										
All	11	6	Good	0.674	0.357	0.688	0.647	0.667	0.666		0.708	0.467
	5	9	Bad	0.643	0.353	0.600	0.643	0.621	0.666		0	0.500
	k = 0.2881		Weighted Avg.	0.645	0.355	0.648	0.645	0.646	0.666		0.388	0.482
Arts	8	9	Good	0.471	0.500	0.533	0.471	0.500	0.492		0.708	0.571
	7	7	Bad	0.500	0.529	0.438	0.500	0.467	0.492		0	0.444
	k = -0.029		Weighted Avg.	0.484	0.513	0.490	0.484	0.485	0.492		0.388	0.514
Background	11	6	Good	0.647	0.714	0.524	0.647	0.579	0.546		0.708	0.516
	10	4	Bad	0.286	0.353	0.400	0.286	0.333	0.546		0	0.516
	k = -0.069		Weighted Avg.	0.484	0.551	0.468	0.484	0.468	0.546		0.388	0.516
Lifestyle	11	6	Good	0.647	0.286	0.733	0.647	0.688	0.706		0.708	0.313
	4	10	Bad	0.714	0.353	0.625	0.714	0.667	0.706		0	0.267
	k = 0.3568		Weighted Avg.	0.677	0.316	0.684	0.677	0.678	0.706		0.388	0.292
Pastime	11	6	Good	0.647	0.214	0.786	0.647	0.710	0.607		0.708	0.733
	3	11	Bad	0.786	0.353	0.647	0.786	0.710	0.607		0	0.750
	k = 0.4247		Weighted Avg.	0.710	0.277	0.723	0.710	0.710	0.607		0.388	0.741
Personality	11	6	Good	0.647	0.857	0.478	0.647	0.550	0.347		0.708	0.647
	12	2	Bad	0.143	0.353	0.250	0.143	0.182	0.347		0	0.571
	k = -0.2183		Weighted Avg.	0.419	0.629	0.375	0.419	0.384	0.347		0.388	0.613
Sports	16	1	Good	0.941	1	0.533	0.941	0.681	0.294		0.708	0.647
	14	0	Bad	0	0.059	0	0	0	0.294		0	0.571
	k = -0.0641		Weighted Avg.	0.516	0.575	0.292	0.516	0.373	0.294		0.388	0.613
Wrapper	13	4	Good	0.765	0.214	0.813	0.765	0.788	0.754		0.708	0.733
	3	11	Bad	0.786	0.235	0.733	0.786	0.759	0.754		0	0.750
	k = 0.547		Weighted Avg.	0.774	0.224	0.777	0.774	0.775	0.754		0.388	0.741
Manual	10	7	Good	0.588	0.071	0.909	0.588	0.714	0.664		0.708	0.714
	1	13	Bad	0.929	0.412	0.650	0.929	0.765	0.664		0	0.765
	k = 0.498		Weighted Avg.	0.742	0.225	0.792	0.742	0.737	0.664		0.388	0.737

Table 31: Results of Experiment 6

Results are in Table 31. Classification trainings using Wrapper-selected features showed the highest F of Good at 0.788 and the highest F of the weighted average between two classes at 0.775. The second highest was using Manually-selected features: F-Measure of the Good class was 0.714 and F-Measure for the weighted average between two classes was 0.737. Trainings from attribute subsets Lifestyle and Pastimes presented fair results although Recall of Good was weaker than that of Bad. For the Pastime feature set, OneR returned better F-Measure for Good.

J48:

pastime_board_scrabble = TRUE: good (13.0/2.0)

pastime_board_scrabble = FALSE

| background_use_of_ja_length = 16 to 20 years: bad (0.0)

| background_use_of_ja_length = lifetime: bad (14.0/3.0)

| background_use_of_ja_length = 1 to 5 years: bad (1.0)

| background_use_of_ja_length = More than 20 years: good (3.0)

OneR:

pastime_board_scrabble:

TRUE → good

FALSE → bad

Figure 11: J48 Decision Tree and OneR Selected Feature for Wrapper Selection in Experiment 6

J48:

pastime_view_opera = FALSE

| backgrond_lang_qual_ja = FALSE: bad (20.0/6.0)

| backgrond_lang_qual_ja = TRUE: good (5.0)

pastime_view_opera = TRUE: good (6.0)

OneR:

music_preference_eu:

Tchaikovsky - Piano Concerto No.1 → bad

Not familiar with any of these → bad

Fauré - Élégie for Cello → bad

Bach - Magnificat → bad

Beethoven - 5th Symphony → good

Figure 12: J48 Decision Tree and OneR Selected Feature for Manual Selection in Experiment 6

J48:

pastime_view_opera = FALSE

| newspaper_read_dailyteleg = FALSE

| | newspaper_read_mainichi = TRUE: bad (5.0)

| | newspaper_read_mainichi = FALSE

| | | pastime_view_theatre = FALSE: good (13.0/5.0)

| | | pastime_view_theatre = TRUE: bad (4.0)

| newspaper_read_dailyteleg = TRUE: good (3.0)

pastime_view_opera = TRUE: good (6.0)

OneR:

pastime_view_opera:

FALSE → bad

TRUE → good

Figure 13: J48 Decision Tree and OneR Selected Feature for Lifestyle Features in Experiment 6

J48:

```
pastime_card_bridge = FALSE
|   pastime_board_scrabble = FALSE: bad (15.0/3.0)
|   pastime_board_scrabble = TRUE
|   |   pastime_like_cryptic_puzzle = No: good (8.0)
|   |   pastime_like_cryptic_puzzle = Yes: bad (3.0/1.0)
pastime_card_bridge = TRUE: good (5.0)
```

OneR:

```
pastime_board_scrabble:
FALSE  → bad
TRUE   → good
```

Figure 14: J48 Decision Tree and OneR Selected Feature for Pastime Features in Experiment 6

Based on accuracy in Lifestyle, Pastime, Wrapper, and Manual, J48 scored 0.677, 0.710, 0.774, and 0.742 (Table 32); OneR scored 0.290, 0.742, 0.742, and 0.742. OneR in lifestyle is too low. For Pastime, OneR is slightly better, for Wrapper, J48 is better and for Manual they are equal.

For Lifestyle features (Figure 13), neither J48 nor OneR was statistically significant. For Pastime (Figure 14), both J48 and OneR were significant (Table 32). The most important J48 features were pastime_card_bridge, pastime_board_scrabble, and pastime_like_cryptic_puzzle. For OneR, the selected feature was pastime_board_scrabble. Thus these games are important to correct classification, and Scrabble is the one which is shared by both classification algorithms.

For Wrapper features (Figure 11), the performance of both algorithms was significant and the J48 accuracy of 0.774 was the highest of all our experiments. Important J48 features were pastime_board_scrabble and background_use_of_ja_length. Once again, OneR selected pastime_board_scrabble. As Table 33 shows, Scrabble is not correlated to Education (see Section 8.2).

For Manual features (Figure 12), both algorithms were once again significant. J48 selected pastime_view_opera and background_lang_qual_ja while OneR chose music_preference. Both algorithms achieved the same accuracy, 0.742 (Table 32). These results suggest that knowledge of classical music can indicate a good translator. Once again, Opera is not correlated to Education (Section 8.2, Table 33).

8. Validation of Results

8.1 Statistical Significance

We have presented the results of the experiments mainly in terms of F-Measure. However, even a high F-score does not exclude the possibility that results have occurred by chance. We therefore carried out further tests on our results.

Most of the experiments are concerned with binary classification: a translator can either be classified as Good or Bad. The aim of a classifier is to assign all bad translators to the Bad class and all good translators to the Good class. If we have almost the same number of good and bad translators in the participant pool, we can assume a binomial distribution. Hence, we inspect the cumulative probability distribution of the binomial (Howell 2007):

$p = 1 - binomdist(X, N, P)$

where X is the number of correct predictions, N is the number of participants, and P is the probability of success in a particular trial. In each binary classification trial, one translator is assigned either to Good or to Bad. The probability of success in this trial, P, is 0.5, independent of the number of Good

and Bad translators in the participant pool. For the three-way classifications (all of Experiment 3, Experiment 4 Manual selection balanced class, Wrapper selection balanced class, and Wrapper selection background balanced class) the principle is the same, except that *P* is 0.333.

The results of this test are shown in Table 32 for all the experiments. However, not all would accept that the Good and Bad classes (or Good, Medium and Bad) are sufficiently balanced to allow this test. We therefore carried out a second significance test using the method of bootstrapping (Mooney and Duval 1993; Billinger et al. 2012). In any experiment, there were *N* participants; in Experiments 1-4, *N* was 82, in Experiment 5, *N* was 47, and in Experiment 6, *N* was 31. By the bootstrapping method, a random classifier first assigns a result, either Good or Bad (in the case of a binary classification) to each of the *N* translators in the training set. The accuracy of this is then calculated:

Accuracy = (No. true positives + No. true negatives) / N

This accuracy figure, determined at random as above, is saved and the entire procedure is repeated 4,000 times - 4,000 is considered sufficient to establish significance at the 0.05 level. The result is a list of 4,000 accuracy figures which are in a distribution which approximates to that of the actual experimental data. We now inspect this distribution at the 97.5% level to find the value below which 97.5% of the figures in it lie. Note that this corresponds to the two-tailed form which is the more stringent test. To be significant at the 0.05 level, the accuracy result for a particular classification must be greater than this figure.

When using this method for a three-way classification (in Experiment 3 and in the balanced classes of Experiment 4), the principle is the same except for the calculation of accuracy:

Accuracy = (No. true Good + No. true Medium + No. true Bad) / N

The results of the bootstrap method are also in Table 32. In the body of this article and in our Conclusions (Section 9), we refer to a result as significant only if it passed both the Binomial test and the Bootstrap test. As can be seen in the Table, every experiment included several such instances.

Exp	Features	Tbl	J48 #corr	OneR #corr	Bootstr Percentile	J48 Acc	J48 Binom	J48 Binom Signif?	J48 Bootstr Signif?	OneR Acc	OneR Binom	OneR Binom Signif?	OneR Bootstr Signif?
1	All	25	53	67	0.610	0.646	0.003	Y	Y	0.817	0.000	Y	Y
1	Arts	25	50	50	0.610	0.610	0.018	Y	N	0.610	0.018	Y	N
1	Background	25	60	67	0.610	0.732	0.000	Y	Y	0.817	0.000	Y	Y
1	Lifestyle	25	57	64	0.610	0.695	0.000	Y	Y	0.780	0.000	Y	Y
1	Pastime	25	56	56	0.610	0.683	0.000	Y	Y	0.683	0.000	Y	Y
1	Personality	25	59	57	0.610	0.720	0.000	Y	Y	0.695	0.000	Y	Y
1	Sports	25	59	58	0.610	0.720	0.000	Y	Y	0.707	0.000	Y	Y
2	All	26	39	56	0.610	0.476	0.630	N	N	0.683	0.000	Y	Y
2	Arts	26	48	38	0.610	0.585	0.049	Y	N	0.463	0.709	N	N
2	Background	26	55	56	0.610	0.671	0.001	Y	Y	0.683	0.000	Y	Y
2	Lifestyle	26	51	57	0.610	0.622	0.010	Y	Y	0.695	0.000	Y	Y
2	Pastime	26	39	39	0.610	0.476	0.630	N	N	0.476	0.630	N	N
2	Personality	26	44	42	0.610	0.537	0.220	N	N	0.512	0.370	N	N
2	Sports	26	40	49	0.610	0.488	0.544	N	N	0.598	0.030	Y	N
3	All	27	36	54	0.440	0.439	0.017	Y	N	0.659	0.000	Y	Y
3	Arts	27	41	36	0.440	0.500	0.001	Y	Y	0.439	0.017	Y	N
3	Background	27	50	54	0.440	0.610	0.000	Y	Y	0.659	0.000	Y	Y
3	Lifestyle	27	45	44	0.440	0.549	0.000	Y	Y	0.537	0.000	Y	Y
3	Pastime	27	33	30	0.440	0.402	0.076	N	N	0.366	0.227	N	N
3	Personality	27	28	28	0.440	0.341	0.388	N	N	0.341	0.388	N	N
3	Sports	27	37	36	0.440	0.451	0.010	Y	Y	0.439	0.017	Y	N
4A	Manual STB	28	60	60	0.610	0.732	0.000	Y	Y	0.732	0.000	Y	Y
4A	Wrapper STB	28	62	59	0.610	0.756	0.000	Y	Y	0.720	0.000	Y	Y
4A	WrBGD STB	28	61	67	0.610	0.744	0.000	Y	Y	0.817	0.000	Y	Y
4B	Manual STG	28	57	55	0.610	0.695	0.000	Y	Y	0.671	0.001	Y	Y
4B	Wrapper STG	28	55	59	0.610	0.671	0.001	Y	Y	0.720	0.000	Y	Y
4B	WrBGD STG	28	57	59	0.610	0.695	0.000	Y	Y	0.720	0.000	y	Y
4C	Manual Bal.	28	54	39	0.440	0.659	0.000	Y	Y	0.476	0.003	Y	Y
4C	Wrapper Bal.	28	55	54	0.440	0.671	0.000	Y	Y	0.659	0.000	Y	Y
4C	WrBGD Bal.	28	52	54	0.440	0.634	0.000	Y	Y	0.659	0.000	Y	Y
5A	All STB	29	25	30	0.640	0.532	0.280	N	N	0.638	0.020	Y	N
5A	Arts STB	29	20	22	0.640	0.426	0.809	N	N	0.468	0.615	N	N
5A	Backgnd STB	29	28	33	0.640	0.596	0.072	N	N	0.702	0.002	Y	Y
5A	Lifestyle STB	29	28	31	0.640	0.596	0.072	N	N	0.660	0.009	Y	Y
5A	Pastime STB	29	31	27	0.640	0.660	0.009	Y	Y	0.574	0.121	N	N
5A	Personlty STB	29	31	23	0.640	0.660	0.009	Y	Y	0.489	0.500	N	N
5A	Sports STB	29	21	24	0.640	0.447	0.720	N	N	0.511	0.385	N	N
5A	Wrapper STB	29	34	35	0.640	0.723	0.001	Y	Y	0.745	0.000	Y	Y
5A	Manual STB	29	37	25	0.640	0.787	0.000	Y	Y	0.532	0.280	N	N
5B	All STG	30	32	30	0.640	0.681	0.004	Y	Y	0.638	0.020	Y	N
5B	Arts STG	30	35	33	0.640	0.745	0.000	Y	Y	0.702	0.002	Y	Y
5B	Backgnd STG	30	36	36	0.640	0.766	0.000	Y	Y	0.766	0.000	Y	Y
5B	Lifestyle STG	30	33	35	0.640	0.702	0.002	Y	Y	0.745	0.000	Y	Y
5B	Pastime STG	30	35	37	0.640	0.745	0.000	Y	Y	0.787	0.000	Y	Y
5B	Personlty STG	30	36	37	0.640	0.766	0.000	Y	Y	0.787	0.000	Y	Y
5B	Sports STG	30	35	38	0.640	0.745	0.000	Y	Y	0.809	0.000	Y	Y
5B	Wrapper STG	30	34	37	0.640	0.723	0.001	Y	Y	0.787	0.000	Y	Y
5B	Manual STG	30	35	36	0.640	0.745	0.000	Y	Y	0.766	0.000	Y	Y
6	All	31	20	15	0.680	0.645	0.035	Y	N	0.484	0.500	N	N
6	Arts	31	15	18	0.680	0.484	0.500	N	N	0.581	0.141	N	N
6	Background	31	15	16	0.680	0.484	0.500	N	N	0.516	0.360	N	N
6	Lifestyle	31	21	9	0.680	0.677	0.015	Y	N	0.290	0.985	N	N
6	Pastime	31	22	23	0.680	0.710	0.005	Y	Y	0.742	0.002	Y	Y
6	Personality	31	13	11	0.680	0.419	0.763	N	N	0.355	0.925	N	N
6	Sports	31	16	12	0.680	0.516	0.360	N	N	0.387	0.859	N	N
6	Wrapper	31	24	23	0.680	0.774	0.000	Y	Y	0.742	0.002	Y	Y
6	Manual	31	23	23	0.680	0.742	0.002	Y	Y	0.742	0.002	Y	Y

Table 32: Statistical Significance. The columns show the experiment number, the features involved, the results table (see earlier in paper), the number correct for J48 and OneR, Bootstrap percentile, J48 accuracy, Binomial, Binomial significance, Bootstrap significance, OneR accuracy, Binomial, Binomial significance, Bootstrap significance. Binomial values less then 0.001 are shown as 0.000.

8.2 Correlation of Key Features

One of the findings of our study was that interest in Opera and enjoyment of the word game Scrabble can indicate a good translator. However, it could be argued that highly educated translators are likely to have interests of this kind, and that a person's ability at translation is simply a consequence of their education and not these pastimes. To test this hypothesis, we computed Pearson's correlation between each of three educational features and the two pastimes, as shown in Table 33. These figures are across the entire population of 82 participants in our study.

Correlations between the three Education features and Scrabble were very low; the maximum was 0.145 between M.A./M.Sc. and Scrabble. The figures for Opera were slightly higher for B.A./B.Sc. (0.215) and M.A./M.Sc. (0.278). However, these are still very weak. Note also that members of the M.A./M.Sc. class were not good at the translation task either. So these results support our hypothesis that qualifications are not the only indicator of a good translator.

Education	Opera	Scrabble	Translation Class Good/Bad
B.A. B.Sc.	0.215	-0.009	0.434
M.A. M.Sc.	0.278	0.145	0.269
Ph.D.	0.048	-0.065	0.076

Table 33: Pearson Correlation between Education Features and Two Pastime Features, Opera and Scrabble, for the 82 Participants

8.3 J48 Cross Validation and Alternative Parameters

We tested another type of cross validation, Leave-One-Out, to see if it returned better results from the rather small number of instances in the dataset. This method leaves one instance out for the validation, uses the rest for the training and averages the results (Witten et al 2011). This method can result in increased accuracy, since almost all of the instances in the dataset can be used for training. We also tested pruning to see if a better accuracy could be gained by adjusting the confidence factor: a smaller value incurs more pruning. We used 'CVParameterSelection' in WEKA to change the confidence parameter from 0.1 to 0.5 by a step of 0.1 (Witten et al 2011). The higher the confidence parameter is, the less pruned. We did not find any improvement by the use of parameter adjustment; the use of Leave-One-Out cross validation resulted in small improvements to Weighted Average F for Manual in Experiment 5, and for Pastime and Wrapper in Experiment 6. However, Ten-Fold cross-validation is normally considered the stronger, so we have restricted our claims to those figures. Table 34 below shows the key findings.

Expt.	Feature	Number of Trans.	Actually Good Trans.	Actually Bad Trans.	Weighted Ave. F using 10-Fold Cross Validation	Weighted Ave. F using Leave-One-Out	Weighted Ave. F using Confidence Parameter Adjustment
5	Personality	47	20	27	0.603	0.603	0.553
5	Manual	47	20	27	0.774	0.799	0.774
6	Lifestyle	31	17	14	0.678	0.677	0.646
6	Pastime	31	17	14	0.710	0.741	0.676
6	Wrapper	31	17	14	0.775	0.807	0.676
6	Manual	31	17	14	0.737	0.701	0.737

Table 34: Cross Validation using Leave-One-Out and Confidence Parameter Adjustment relating to Key Results from Experiments 5 & 6

9. Conclusions

We have explained how we created a dataset, based on 82 participants, which includes 146 attributes and a translation score for each. Based on this, six decision-tree training experiments with variations of classification, participants, feature selection and class skews were conducted using J48 with 10-fold cross validation. Comparisons were also made with OneR. We now summarise the main findings.

If the object is to select a Good translator to carry out some work, then the Precision score on class Good tells you the proportion of candidates selected by the decision tree who are actually good. Over the entire population of translators, the best Precision on class Good using a predefined subset of the features (0.678) was in Experiment 2 (threshold 0.7) using Background features (Table 26). This result was statistically significant ($p < 0.05$, see Table 32). Figure 4 shows the decision tree which only uses University study, English language qualification, and fluency in French. Lifestyle features were close behind (0.667) and the decision tree can be seen in Figure 5. This result was also significant ($p < 0.05$, Table 32). Playing in a rock band, reading various newspapers, and watching Rakugo were taken into account.

Over the entire population, and using Wrapper selection, the best Precision on class Good (0.731) was in Experiment 4 (Wrapper selection Skew to Good, threshold 0.7, Table 28; significant, $p < 0.05$, Table 32). Figure 7 shows the decision tree. Reading the Nikkei, not being a student, and living in a country where their learned language is spoken were important here. Interestingly, not wearing a wristwatch, and going straight or right at a crossroad also participated in the selection.

Note that in both the above cases, we are using the lower threshold of 0.7 (the upper was 0.8) meaning that a Good translator is only 'fairly good'.

Student translators generally fared poorly in our study as our test sentences were far too difficult for them. If students are excluded from the training data, the best Precision on class Good (0.917) was in Experiment 5, Skew to Bad (threshold 0.8) using Manual selection (Table 29, significant, $p < 0.05$, Table 32). This was the best overall result of the study in terms of selecting a Good translator. The decision tree is in Figure 10 and uses just two features, reading the Daily Telegraph and playing the card game Bridge.

Professional translators performed well in the translation task. In Experiment 6, Skewed to Good (threshold 0.8) using Manual selection (Table 31) showed another high Precision (0.909). This is the second best Precision in the entire experiments (significant, $p < 0.05$, Table 32). The decision tree is in Figure 12 and used only two features, going to the Opera as a pastime activity and holding a Japanese language qualification.

The above results relate to success at selecting a good translator from the point-of-view of the employer. However, they do not take into account Good translators who are wrongly classified as Bad and are therefore not selected and neither do they measure the overall accuracy of the classifier. Next we look into the overall abilities of the classifiers based on the weighted F-score calculated by WEKA.

Over the entire population, the best weighted F-score (0.728) using a predefined subset of the features was in Experiment 1 (threshold 0.8, Table 25) using Background features. This was statistically significant ($p < 0.05$, see Table 32). However, Precision of the Good class was weak (0.524). The second best was 0.660, also in Experiment 1, which again had a low Precision on Good (0.417) and this time used Lifestyle features (significant, $p < 0.05$, Table 32). Pastime also scored 0.660 with an even lower Precision on Good of 0.400. However, the third best was 0.656 in Experiment 2 (threshold 0.7, Table 26) which had a Precision on Good of 0.678, once again using Background features (significant, $p < 0.05$, Table 32). The features in question (Figure 4) were university study, English language qualification and fluency in French.

Over the entire population, the best weighted F-score (0.738) using feature selection was in Experiment 4, Wrapper selection Skew to Bad (threshold 0.8, Table 28) and Wrapper selection

Background Skewed to Bad (also threshold 0.8, Table 28). Both were statistically significant ($p < 0.05$, Table 32). However, both showed low Precision on Good (0.600 and 0.550 respectively). The second best was in Experiment 4 with Wrapper selection Skew to Good (0.704, same table) which also had a higher Precision on Good (0.731). This was also significant ($p < 0.05$, Table 32).

Excluding students, the best weighted F-score (0.774) was in Experiment 5, Manual selection Skew to Bad (threshold 0.8, Table 29). This coincides with the highest Precision on Good of 0.917. Reading the Daily Telegraph and playing the card game Bridge are important here (Figure 10). This was significant ($p < 0.05$, Table 32).

When limiting to professional translators in Experiment 6 (threshold 0.8, Table 31), the best weighted F-score was with Wrapper selection (0.775). This was significant ($p < 0.05$, Table 32). This classifier also showed the second highest Precision on Good (0.813). Figure 11 shows the decision tree; playing Scrabble and the length of Japanese language use are important features. The second best F-score (0.737) has the highest Precision on Good (0.909), using Manual feature selection. Once again this was significant ($p < 0.05$, Table 32).

The results indicate that in addition to the usual questions regarding translators' educational and general background, questions regarding their hobbies and reading habits could also be important for selecting Good translators. We conclude that for selecting a Good translator from a general population not limited to professional translators, important questions one could ask are: if they have lived in a country where their learned language is spoken, if they are fluent in French, if they are a student, what they studied at University, if they have an English language qualification, if they read the Nikkei or the Daily Telegraph, and if they play Bridge. However, if the population is limited to professionals, the best translators could be found by asking if they hold a Japanese language qualification, how long they have been using Japanese, if they go to the Opera and if they play Scrabble. A translator may share some traits with performers or painters; they absorb external information and express it eloquently for us. They do this by being persistent and meticulous in details.

The results of this work are clearly impeded by the small sample size comprising just 82 translators which, combined with the large set of features used, makes it very difficult to train a system properly. Nevertheless, it was not easy to recruit even those 82, as this involved extensive campaigning in America, Australia, Canada, Ireland, Japan, and the United Kingdom over many months. In considering which features were found important in the various decision trees, we have to bear this in mind; the trees are interesting in suggesting the relevance of certain facts about a translator (and many of the results were statistically significant as already reported), but this study alone cannot determine the predictive power of each.

In future, findings could be refined further by increasing the sample size and adding further features. Also, while we arbitrarily set the quality threshold of a Good translation, it would be useful to find out what is the accepted dividing point between Good and Bad translations.

Acknowledgement

We would like to thank the 82 participants and the four evaluators in this study. This work could not have been achieved without their generous contributions. We are also indebted to Rebecca Bourke for her help. Finally, we thank the anonymous reviewers for their careful reading and detailed comments.

References

Alpaydin, E. (2004) *Introduction to Machine Learning* Cambridge, MA: MIT Press.

Avramidis, E., Popovic, M., Vilar, D. and Burchardt, A. (2011) Evaluate with Confidence Estimation: Machine Ranking of Translation Outputs Using Grammatical Features, in the *6th Workshop on Statistical Machine Translation, Edinburgh, Scotland, UK, 30-31, July, Association for Computational Linguistics,* 65-70.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Ann Arbor, Michigan.*

Biel, Ł. (2011) Training Translators or Translation Service Providers? EN 15038:2006 standard of translation services and its training implications, *The Journal of Specialised Translation*, (16), 61-76.

Billinger, M., Daly, I., Kaiser, V., Jin, J., Allison, B. Z., Müller-Putz, G. R. and Brunner, C. (2012) Is it significant? Guidelines for reporting BCI performance, In B. Allison et al. (eds) *Towards Practical Brain-Computer Interfaces*, 333-354, Berlin, Germany: Springer.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A. and Ueffing, N. (2004) Confidence Estimation for Machine Translation, in *Proceedings of the 20th Conference on Computational Linguistics, Geneva,* 315-321.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L. and Tamchyna, A. (2014) Findings of the 2014 Workshop on Statistical Machine Translation. *Proceedings of the Ninth Workshop on Statistical Machine Translation, 12–58, Baltimore, Maryland, USA, June 26–27.*

Brau, M. (2014) The FBI Develops a Translation Aptitude Test (TAT), *Interagency Language Round Table Plenary Session*, Language Testing and Assessment Unit, National Foreign Language Center (NFLC), College Park, MD.

Callison-Burch, C. (2009) Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk, in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 286-295.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Shroeder, J. (2007) Meta - Evaluation of Machine Translation, in *Proceedings of ACL Workshop on Statistical Machine Translation*, 136-158.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M. and Zaidan, O. F. (2010) 'Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation', in *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, Association for Computational Linguistics*, 17-53.

Doddington, G. (2002) Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. *Proceedings of the second international conference on Human Language Technology Research*, 138-145.

Fomicheva, M., Bel, N., Specia, L., da Cunha, I. and Malinovskiy, A. (2016) CobaltF: A Fluent Metric for MT Evaluation. *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, 483-490, Berlin, Germany, August 11-12.

Gile, D. (1995) *Basic Concepts and Models for Interpreter and Translator Training*, Amsterdam: John Benjamins Publishing Co.

Graham, Y., Mathur, N. and Baldwin, T. (2015) Accurate Evaluation of Segment-level Machine Translation Metrics, *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, Denver, Colorado, May 31 - June 5,* 1183-1191.

Hayes, A. F. and Krippendorff, K. (2007) Answering the Call for a Standard Reliability Measure for Coding Data, *Communication Methods and Measures*, 1(1), 77-88.

Howell, D. C. (2007) *Statistical methods for psychology* (6th ed.). Belmont, Calif.: Thomson.

Hubscher-Davidson, S. E. (2009) Personal Diversity and Diverse Personalities in Translation: A Study of Individual Differences, *Perspectives: Studies in Translatology*, 17(3), 175-192.

Koehn, P. (2004) Statistical Significance Tests for Machine Translation Evaluation. *Proceedings of the Empirical Methods in Natural Language Processing Conference, Barcelona, Spain.*

Koehn, P. (2010) *Statistical Machine Translation*, UK: Cambridge University Press.

Krippendorff, K. (2011) *Computing Krippendorff's Alpha-Reliability*, available: Retrieved from http://repository.upenn.edu/asc_papers/43 [accessed 16/April/2015].

Leusch, G., Ueffing, N. and Ney, H. (2006) CDER: Efficient MT Evaluation using Block Movements. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy*, 241–248.

Mooney, C. Z. and Duval, R. D. (1993) *Bootstrapping: A Nonparametric Approach to Statistical Inference*, London, UK: Sage Publications.

Myers, I. B. and Myers, B. (1995) *Gifts differing. Understanding personality type*. Mountain View, CA: Davies-Black Publishing.

Nagao, M., Tsujii, J. and Nakamura, J. (1985) The Japanese Government Project for Machine Translation, *Computational Linguistics*, 11(2-3), 91-110.

Napthine, A. (1983) Training of Translators, in Picken, C., ed. *The Translator's Handbook*, London: Aslib, 21-32.

Nida, E. (1981) Translators are Born Not Made, *Practical Papers for The Bible Translator*, 32(4), 401-405.

O'Brien, S. (2013) The Borrowers: Researching the Cognitive Aspects of Translation, *Target - International Journal of Translation Studies*, 25(1), 5-17.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation, in the *40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 311-318*.

Popović, M. and Ney, H. (2007) Word Error Rates: Decomposition over POS Classes and Applications for Error Analysis, in the *Second Workshop on Statistical Machine Translation, Prague, Association for Computational Linguistics, 48-55.*

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, Inc.

Rahemi, E. F., Jufri. and Ardi, H. (2013) The Correlation Between Reading Comprehension and Translation Ability: A Correlational Study on Forth Year Students at English Department of UNP, *Journal of English Language Teaching*, 1(2), 178-186.

Sakaguchi, K., Post, M. and Van Durme, B. (2014) Efficient elicitation of annotations for human evaluation of machine translation. *Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland.*

Schopp, J. F. (2007) The European Translation Standard EN 15038 and it's Terminology - A Mirror of Missing Professionalism?. *ELETO-6th Conference Hellenic Language and Terminology, Athens, Greece.*

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006) A Study of Translation Edit Rate with Targeted Human Annotation, *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006), Cambridge, Massachusetts.*

Specia, L., Cancedda, N., Dymentman, M., Turchi, M. and Cristianini, N. (2009) Estimating the Sentence-Level Quality of Machine Translation Systems, in *Proceedings of the 13th Annual Conference of the EAMT, Barcelona*, 28-35.

Specia, L., Raj, D. and Turchi, M. (2010) Machine Translation Evaluation versus Quality Estimation, *Machine Translation*, 24, 39-50.

Specia, L. and Shah, K. (2014) Predicting Human Translation Quality, in Al-Onarizan, Y. and Simard, M., eds., *The 11th Conference of the Association for Machine Translation in the Americas, Vancouver*, 288-300.

Sumita, E., Sasaki, Y. and Yamamoto, S. (2005) The Forefront of Evaluation Methods for Machine Translation Systems, *IPSJ Magazine*, 552-557.

SurveyMonkey® (2014) *SurveyMonkey®*, [online], available: http:// www.surveymonkey.net/ ?ut_source=header [accessed 16/April/2015].

Suzuki, A. (1988) Aptitude of Translators and Interpreters, *Meta: Translators' Journal,* 33(1), 108-114.

Tavakoli, M., Shafiei, S. and Hatam, A. H. (2012) The Relationship between Translation Tests and Reading Comprehension: A Case of Iranian University Students, *Iranian Journal of Applied Language Studies*, 4(1), 193-211.

The European Commission (2015) *Translation*, [online], available: http://ec.europa.eu/ dgs/translation/programmes/emt/index_en.htm [accessed 16/April/2015].

The House of Representatives of Japan (2014) *The Transcript of Meetings*, [online], available: http://www.shugiin.go.jp/internet/itdb_kaigiroku.nsf/html/kaigiroku/kaigi_l.htm [accessed 16/ April/2015].

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997) Accelerated DP based search for statistical translation. In Kokkinakis, G., Fakotakis, N., and Dermatas, E., editors, *Proceedings of the Fifth European Conference on Speech Communication and Technology, 2667–2670, Rhodes, Greece. International Speech Communication Association.*

Turian, J. P., Shen, L. and Melamed, I. D. (2003) Evaluation of Machine Translation and its Evaluation, in *Proceedings of MT Summit IX, New Orleans, LA.*, 386-393.

Verrinder, J. (1983) Who are the translators? in Picken, C., ed. *The Translator's Handbook*, London: Aslib, 33-37.

Voorhees, E. M. (2003) Overview of the TREC 2003 Question Answering Track, in *TREC 2004*, NIST, Gaithersburg, Md. USA.

Witten, I. H. (2013) *Data Mining with Weka (3.5: Pruning decision trees)*, [online], available: http://www.youtube.com/watch?v=ncR_6UsuggY [accessed 16/April/2015].

Witten, I. H., Frank, E. and Hall, M. A. (2011) *Data Mining – Practical Machine Learning Tools and Techniques*, 3rd ed., USA: Morgan Kaufmann.

Zaidan, O. F. and Callison-Burch, C. (2011) Crowdsourcing Translation: Professional Quality from Non-Professionals, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, Association for Computational Linguistics*, 1220-1229.