Analysing Lexical Consistency in Translation

Liane Guillou
School of Informatics
University of Edinburgh
Scotland, United Kingdom
L.K.Guillou@sms.ed.ac.uk

Abstract

A number of approaches have been taken to improve lexical consistency in Statistical Machine Translation. However, little has been written on the subject of where and when to encourage consistency. I present an analysis of human authored translations, focussing on words belonging to different parts-of-speech across a number of different genres.

1 Introduction

Writers are often given mixed messages with respect to word choice. On one hand they are encouraged to vary their use of words (in essay writing): "It is also important that the words you use are varied, so that you aren't using the same words again and again."[1]. On the other hand they are encouraged to use the same words (only changing the determiner) when referring to the same entity a second time (in technical writing): "The first time a single countable noun is introduced, use *a*. Thereafter, when referring to that same item, use *the*." [2].

Halliday and Hassan (1976) showed that well-written documents exhibit lexical cohesion in terms of what they call *reiteration* and *collocation*. Reiteration is achieved via repetition as well as the use of synonyms and hypernyms. A collocation is a sequence of words / terms that co-occur regularly in text. Examples of collocated pairs of words include "fast food", "bright idea" and "nuclear family". Any source language document will therefore contain repeated instances of the same words or *lemmas* (morphological variants of the same words). This repeated use of words and lemmas is known as *lexical consistency* and the instances can be grouped together to form *lexical chains* (Morris and Hirst, 1991). Lexical chains were proposed by Lotfipour-Saedi (1997) as one feature of a text via which translational equivalence between source and target could be measured.

While Statistical Machine Translation (SMT) has gone from ignoring these properties of discourse by translating sentences independently, to trying to impose *lexical consistency* at a universal level, both approaches have given little consideration to what might be standard practice among human translators.

In order to discover what the standard practice might be, and thus what an SMT system might better aim to achieve, I have carried out a detailed analysis of lexical consistency in human translation. For comparison, I also present an analysis of translations produced by an SMT system. I have considered a variety of genres, as genre correlates with the *function* of a text, which in turn predicts its important elements. A preliminary conclusion of this analysis is that human translators use lexical consistency to support what is important in a text.

2 Related Work

2.1 Unique Terms and Lexical Consistency

Intuitively, it seems obvious that specialised, "semantically heavy" words like "genome" and "hypochondria" will only have a single exact translation into any given target language, and as such will tend to be translated with greater consistency than semantically "light" words. Melamed (1997) showed that this intuition could be quantified using the concept of *entropy*, which the

[1] Purdue University, Online Writing Lab: http://owl.english.purdue.edu/engagement/index.php?category_id=2&sub_category_id=2&article_id=66. Accessed 21/04/2013

[2] Monash University, Language and Learning Online: http://monash.edu.au/lls/llonline/grammar/engineering/articles/6.xml. Accessed 21/04/2013

Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT), pages 10–18,
Sofia, Bulgaria, August 9, 2013.

author uses over a large corpus to show what words and what parts-of-speech are more likely to be translated consistently than others. However, Melamed's analysis ignores any segmentation of the corpus by document, topic, speaker/writer or translator, considering only overall translational distributions. It is therefore similar to that which can be gleaned from the phrase table in a modern SMT system.

2.2 Enforcing and Encouraging Consistency

A number of approaches have been taken to both encourage and enforce lexical consistency in SMT. These range from the cache-based model approaches of Tiedemann (2010a; 2010b) and Gong et al. (2011), to the post-editing approach of Xiao et al. (2011) and discriminative learning approach of Ma et al. (2011) and He et al. (2011).

Carpuat (2009) and Ture et al. (2012) suggested that the *one sense per discourse* constraint (Gale et al., 1992) might apply as well to *one sense per translation*. Both demonstrated that exploiting this constraint in SMT led to better quality translations. Ture et al. (2012) encourage consistency themselves using soft constraints implemented as additional features in a hierarchical phrase-based translation model.

What has not been adequately addressed in the available MT literature is *where* and *when* lexical consistency is desirable in translation.

2.3 Measuring Consistency

In contrast with *entropy* following from lexical properties of words (i.e. how many senses a word has, and how many different possible ways there are of translating each sense in a given target language), as explored in (Melamed, 1997), Itagaki et al (2007) developed a way to measure the terminological consistency of a *single* document. They define *consistency* as a measure of the number of translation variations for a term and the frequency for each variation. They adapted the Herfindahl-Hirschman Index (HHI) measure, typically used to measure market concentration, to measure the consistency of a single term in a single document. HHI is defined as:

$$HHI = \sum_{i=1}^{n} s_i^2$$

Where *i* ranges over the *n* different ways that the given term has been translated in the document, and s_i is the ratio of the number of times the term has been translated as *i* to the number of times it has been translated. The lower the index, the more variation there is in translation of the term, i.e. the less consistent the translation. The maximum index is 10,000 (or 1 using the normalised scale) for a completely consistent translation.

HHI is best illustrated with examples of distributions over a *single document*. An English word with two French translations that are observed with equal frequency will receive a score of: $0.50^2 + 0.50^2 = 0.5$. A different English word with two French translations observed 80% and 20% of the time will receive a score of: $0.90^2 + 0.10^2 = 0.82$ representing a more consistent translation of the English word. When the number of possible French translations increases, the HHI score will likely decrease unless one translation is much more frequent - see previous example. An English word with three translations observed with equal frequency (33.3% each) will have a score of: $0.33^2 + 0.33^2 + 0.33^2 = 0.33$ representing a word that is translated with lower consistency.

Itagaki et al. incorporate these HHI scores (one score per term, per document) in a wider calculation that measures inter-document consistency of a set of documents that all use the same term. As the analyses presented in this paper are concerned with single documents and their translations, the per term, per document HHI scores are sufficient.

3 Methodology

This section describes analyses of manual (human) translation and automated translation (by a phrase-based SMT system). The data used is described in Section 3.1 and the methods for analysing consistency in human and automated translation are described in Sections 3.2 and 3.3.

3.1 Data

As the focus of the analysis is lexical consistency, it was important to select texts that were written/translated by the same author. The typical corpora used in training SMT systems were dismissed; Europarl as speakers change frequently and news-crawl as the articles are typically too short to exhibit much lexical repetition. Instead I selected the INTERSECT corpus (Salkie, 2010) which contains a collection of sentence-aligned parallel texts from different genres. From this corpus I extracted a number of texts from the English-

Title	Genre	Sentences	Words		En POS Count			Fr POS Count		
			En	Fr	N	A	V	N	A	V
English Source										
Xerox ScanWorx Manual	Instructions	2,573	38,698	44,841	14,060	2,308	6,555	15,206	2,528	8,822
On the Origin of Species	Natural Science	1,702	62,454	68,016	13,774	6,868	9,857	17,452	6,291	12,895
Dracula Ch. 1-2	Novel	584	11,209	10,840	2,147	817	2,110	2,659	745	2,336
The Invisible Man Ch. 1-4	Novel	504	7,578	7,924	1,845	442	1,471	2,118	472	1,720
French Source										
Nuclear Testing	Public Info	613	13,127	13,563	3,918	1,412	1,808	4,261	1,344	2,253
French Revolution to 1945	Public Info	1530	34,038	33,187	11,217	3,119	4,279	11,008	3,025	4,632
The Immoralist	Novel	1,377	29,323	24,942	5,299	2,049	5,888	5,813	1,513	6,138
News article 1	News	126	1,757	1,751	549	122	284	558	115	324
News article 2	News	126	2,306	2,254	590	150	430	673	125	459
News article 3	News	85	1,891	1,756	501	183	332	534	122	332
News article 4	News	97	2,236	1,974	641	157	367	609	120	356

Table 1: Documents taken from the English (En) - French (Fr) section of the INTERSECT corpus.

French collection (Table 1). The frequencies for nouns (N), adjectives (A) and verbs (V) in this table were extracted automatically using the TreeTagger tool (Schmid, 1994).

Word alignments for the parallel documents were computed using Giza++ (Och and Ney, 2003) run in both directions. In order to improve the robustness of the word alignments the documents were concatenated into a single file, together with English-French parallel data from the Europarl corpus (Koehn, 2005). The word alignments for the relevant documents were then extracted from the symmetrised alignment file.

3.2 Consistency in Human Translation

The motivation for this analysis was to assess the extent to which a human translator maintained lexical consistency when translating a document. In other words, in those places where the author of a source document makes consistent lexical choices, do human translators do so as well? And if they do, should we aim for the same in SMT?

For each document, the English and French parallel texts were processed using TreeTagger (Schmid, 1994). Using the language in which the document was originally written (its *born language*) as the source language, word alignments were used to identify what each source word aligned to in the (human) translation.

Since I wanted to establish not just the *degree* of consistency, but *where* consistency was being maintained, and because I felt that the Part-of-Speech (POS) tags output by TreeTagger were too fine-grained for this purpose, these tags were mapped to a set of coarse-grained tags. The Universal POS tagset mapping file (Petrov et al., 2011) was used for English and a comparable file was constructed for French. In addition to this, I also sub-divided the coarse-grained verb class into three classes: light verbs (e.g. do, have, make), mid-range verbs (e.g. build, read, speak) and rare-verbs (e.g. revolutionise, obfuscate, perambulate). This was to test the hypothesis that light verbs will exhibit lower levels of consistency than other verbs. A *light verb* is defined a verb with little semantic content of its own that forms a predicate with its argument (usually a noun). For example the verb "do" in "do lunch" or "make" in "make a request". As no predefined lists of light, mid-range and rare verbs are available, these groups were approximated. An English verb's category is determined by its frequency in the British National Corpus (BNC) (Clear, 1993). A verb with a frequency count in the bottom 5% is deemed a rare verb, in the top 5% is deemed a light verb and anything in between, is deemed a mid-range verb. A manual inspection of the resulting category boundaries shows that these thresholds are reasonable. For French, verb frequencies were extracted from the French Treebank (Abeillé et al., 2000).

Herfindahl-Hirschman Index (HHI) (Itagaki et al., 2007) scores were calculated for each surface word (one score per surface word) in the *born language* document. The documents were treated separately, and no inter-document scores are calculated. These scores tell us how consistent the translation is into the target language. For words in the English documents I considered what words *and* lemmas were present in the French translation. Lemmas are included as French verb inflections may otherwise skew the results. For com-

pleteness, lemmas in the English translation of the French documents are also considered.

For each POS category, an average HHI score is calculated by taking the sum of the HHI scores per word and dividing it by the number of words (for that POS category). Only those words that are repeated (i.e. appear more than once in the source document with the same coarse POS category) are considered. (That is, a word that appeared once as a mid-range verb, once as a noun and once as something else, would not be included). A similar average is calculated for lemmas.

HHI scores are normally presented in the range of 0 to 10,000. However, for simplicity, the scores presented in this paper are normalised to between 0 to 1.

3.3 Consistency in Automated Translation

The aim of this analysis was to assess how the consistency in translations produced by an SMT system would compare to those by a human translator. The SMT system was an English-French phrase-based system trained and tuned using (Moses and) Europarl data. Its language model was constructed from the French side of the parallel training corpus. The system was used to translate the *born* English source documents (*Xerox Manual*, *On the Origin of Species*, *Dracula* and *The Invisible Man*). Word alignments and a file containing a list of Out of Vocabulary (OOV) words were also requested from the decoder. Note that *all* of the documents are considered to be "out of domain" with respect to the training data used to build the SMT system.

Using a similar process as described in Section 3.2, but omitting those words that are reported by the decoder as OOV, average HHI scores are calculated for each POS category. OOV words are omitted as these will be "carried through" by the decoder, appearing untranslated in the translation output. They therefore do not say anything about the consistency of the translation.

The other major difference is that HHI scores are calculated only at the word level, not at the lemma level as it is expected that the TreeTagger would perform poorly on SMT output and these errors could lead to misleading results. In all other respects, the process for analysing text is the same as described in Section 3.2.

4 Results

4.1 Consistency in Human Translation

The results are presented in Table 2. Higher (average) HHI scores represent greater consistency.

For both English and French source documents, nouns score highly, suggesting that in general human translators translate nouns rather consistently. However, nouns don't always receive the highest average score. For verbs, the trend is that consistency is irrelevant in translating light verbs, rare verbs tend to be translated with the highest consistency, and mid-range verbs are somewhere in between. This suggests that consistency in the translation of light verbs would be undesirable.

Looking at some of the texts in more detail it may be possible to infer certain qualities of text across different genres.

Novels: In all three texts, (*Dracula*, *The Invisible Man* and *The Immoralist*), nouns receive the highest average HHI score of all the POS categories. An analysis of some of the most frequent (and aligned) nouns in *Dracula* (Table 3) suggests that it is desirable to keep important nouns constant - those that identify characters and other entities central to the story. For example, the *Count* is an important character and is never referred to by any other name/title in the original text. (N.B. "count" is also a mid-range verb, but it is used only as a noun in *Dracula*). The translation to *(le) comte* in French is highly consistent. A similar observation is made for *horses* which are important in the story. Interestingly, the (same) coach *driver* is referred to as *(le) chauffer*, *(le) conducteur* and *(le) cocher* in French:

English: ...and the *driver* said in excellent German
French: Le *conducteur* me dit alors, en excellent allemand

English: Then the *driver* cracked his whip
French: Puis le *chauffeur* fit claquer son fouet

English: When the caleche stopped, the *driver* jumped down
French: La calèche arrêtée, le *cocher* sauta de son siège

This perhaps reflects a stylistic choice made by the translator to vary the terms used to refer to a character of lesser importance. It is worth noting that the English text also contains several instances of "coachman" to refer to the "driver" but the variation is much less compared with the French translation.

Verbs, on the other hand, receive lower (average) HHI scores indicating that this may be an area

Title	Noun	Adj	Verb			
			All	Light	Mid-Range	Rare
English Source						
Xerox ScanWorx Manual	**0.6995**	0.5900	0.5568	0.3256	0.5766	0.6485
Xerox ScanWorx Manual (Lemmas)	**0.7126**	0.7112	0.6612	0.4172	0.6902	0.7086
On the Origin of Species	**0.6109**	0.4390	0.4001	0.2339	0.4140	0.4592
On the Origin of Species (Lemmas)	**0.6417**	0.5722	0.5056	0.3355	0.5273	0.5098
Dracula	**0.6182**	0.4191	0.3631	0.2477	0.4175	0.5000
Dracula (Lemmas)	**0.6294**	0.4979	0.4113	0.2902	0.4711	0.5000
The Invisible Man	**0.6290**	0.5110	0.4159	0.3139	0.4797	0.4219
The Invisible Man (Lemmas)	**0.6275**	0.5743	0.4573	0.3723	0.5121	0.4219
French Source						
Nuclear Testing	0.7388	**0.8079**	0.5616	0.3312	0.5279	0.6228
Nuclear Testing (Lemmas)	0.7521	**0.8209**	0.5972	0.4198	0.5599	0.6584
French Revolution to 1945	0.6346	**0.6587**	0.5054	0.3041	0.4404	0.5521
French Revolution to 1945 (Lemmas)	0.6509	**0.6632**	0.5266	0.3950	0.4710	0.5655
The Immoralist	**0.6807**	0.5732	0.4868	0.3106	0.4524	0.5046
The Immoralist (Lemmas)	**0.7007**	0.5856	0.5142	0.3821	0.4977	0.5236
News article 1	**0.7278**	0.6400	0.5424	0.4336	0.5608	0.5734
News article 1 (Lemmas)	**0.7542**	0.6400	0.5616	0.4943	0.5608	0.5911
News article 2	0.6745	**0.7140**	0.5345	0.3660	0.5395	0.6751
News article 2 (Lemmas)	0.6836	0.7140	0.5717	0.4083	0.5395	**0.7778**
News article 3	0.6991	**0.7986**	0.5024	0.3016	0.5794	0.5988
News article 3 (Lemmas)	0.7121	**0.7986**	0.5869	0.4801	0.6508	0.6204
News article 4	**0.6734**	0.6556	0.5073	0.2408	0.6667	0.6295
News article 4 (Lemmas)	0.6984	0.6333	0.6118	0.3790	0.6667	**0.7545**

Table 2: Human Translation: Average HHI scores for words in the source and their aligned words (and lemmas) in the translations. Scores are provided in the range of 0 to 1 and the highest score for each document is highlighted in bold text. The scores for rare verbs in Dracula and The Invisible Man are the same for words and lemmas. These documents contain very few repeated rare verbs (far fewer than the other English documents) and those that are repeated are very specific and diverse such that no difference is seen between the two distributions.

Noun (word)	HHI score	Count
Count	0.9412	33
driver	0.2985	28
horses	0.9050	20
room	0.4000	20
time	0.1150	20
door	0.5986	17
place	0.6797	16
night	0.4667	15

Table 3: Dracula - most frequent noun words

in which some artistic license may be used.

These findings suggest that when aiming to encourage consistency in the translation of novels, the focus should be on nouns. As for adjectives, less frequent in novels than verbs and nouns (Table 1), further analysis may show whether consistency varies depending on function (e.g. modifier, predicate adjective) or frequency as well. The translation of pronouns also requires investigation.

Natural Science: The natural science text *On the Origin of Species* exhibits a similar pattern of translational consistency to novels. This is perhaps not surprising as 19th century British natural science texts would have had the same middle-class audience as the novels of the same era. The translation of modern scientific texts may or may not follow this pattern.

Instruction Manuals: In the *Xerox Manual* nouns receive the highest average HHI score at the word level. When considering what lemmas the source words align to in the translation, nouns again score the highest, closely followed by adjectives and rare verbs. This overall pattern makes sense as in an instruction manual it is important to identify both the actions and entities involved at each step. Adjectives will help the user correctly identify the intended entities. The word-level HHI scores for the most frequently used (and aligned) rare verbs are given in Table 4.

The verb *process* has several translations in French: *traitement* ("treatment"/"processing"), *traiter* ("process") and *exécuter* ("execute"). (Note that *traitement* is in fact a noun, reflecting a change in the structure of the sentence.) The

Rare Verb (word)	HHI score	Count
process	0.5729	109
previewing	0.5868	33
previewed	0.6399	19
verifying	0.5556	18
formatted	0.3244	15
scans	1.0000	13
formatting	0.4380	11
dithering	0.4380	11

Table 4: Xerox Manual - most frequent rare verb words

Adjective (word)	HHI score	Count
nationale (*national*)	0.8233	75
européenne (*European*)	0.8232	64
économique (*economic*)	0.8575	40
constitutionnel (*constitutional*)	0.9474	37
française (*French*)	0.4288	37
constitutionnelle (*constitutional*)	1.0000	31
français (*French*)	0.7899	26
autres (*other*)	0.8496	25

Table 5: French Revolution to 1945 - most frequent adjective words

resulting translations into French are all clear, so this may simply be a reflection of a difference in terminology between English and French, at least as used by Xerox. For example:

English: *Process* the page and save the output as an image.
French: *Traitement* de la page et sauvegarde de la sortie comme image.

English: Page Settings enable you to describe the pages that the system is about to *process*.
French: Les Instructions de page vous permettent de décrire les pages que le système va *traiter*.

English: Load Verification Data, Loads a named verification data file to *process* a job.
French: Charger données de vérification, Charge un fichier nommé de données de vérification pour *exécuter* une tâche.

What is also interesting is that in the English text, the word *process* is used as both a noun and a rare verb. However, it is translated more consistently when used as a verb (HHI: 0.5729) compared with its use as a noun (HHI: 0.2576).

In this genre, accuracy and readability are important and it is acceptable to produce a "repetitive" or "boring" text. It may, therefore, be appropriate to encourage translational consistency of nouns, rare verbs and adjectives in instructions. Unlike with novels, it would make sense that *all* entities in an instruction manual are of importance.

Public Information: In the *French Revolution to 1945* and *Nuclear Testing* documents, adjectives score highest, followed by nouns. Word-level HHI scores for the most frequent (and aligned) adjectives in the *French Revolution to 1945* document are presented in Table 5.

Using a manual inspection of those nouns that appear next to (i.e. directly after) the adjective in French, the possibility that these nouns were *semantically light* was explored. Focussing on the English translation, WordNet (Miller, 1995) was used to ascertain the distance of the noun from the root of the relevant hierarchy. The assumption is the semantically light nouns appear closer to the root than other nouns. For all 82,115 noun synsets in WordNet, the average minimum and maximum depths to the root are 7.25 and 7.70 respectively.

Taking the adjective *economic* (*économique* in French) in the *French Revolution to 1945* document as an example, the nouns it is paired with (e.g. expansion, cooperation, development, action, council, etc.) typically have depths below the average and therefore could be considered semantically light. The adjectives used in the text include *constitutionnel / constitutionnelle* ("constitutional"), *économique* ("economic") and *nationale* ("national"). These words are rather specific (or "semantically heavy"), so there may be few alternative valid translations to choose from. This is supported by Melamed's (1997) notion of semantic entropy, in which more specific words receive lower entropy scores, reflecting greater consistency in translation. For texts of this genre, it may be appropriate to encourage the consistent translation of adjectives and nouns, allowing for more freedom in the translation of verbs.

News Articles: The pattern for news articles is a little less predictable, although a similar pattern (to other document types) can be seen for light, mid-range and rare verbs. This may be due to the short length of the texts (circa 2,000 words) which may not be sufficient to establish a stable pattern. Or it may be that there are different writing styles within the news genre dependent on the type or subject of the "story".

4.2 Consistency in Automated Translation

The results of a similar analysis of translational consistency in phrase-based SMT are presented in Table 6. Overall, consistency is much higher than in translations produced by human translators. But what does this mean? Is the problem of consistency in SMT non-existent? In short, no; there are

POS Category	Xerox Manual		Origin of Species		Dracula		The Invisible Man	
	Automated	Human	Automated	Human	Automated	Human	Automated	Human
Noun	0.8502	0.6995	0.8481	0.6109	0.8318	0.6182	0.8308	0.6290
Adj	0.6871	0.5900	0.6333	0.4390	0.6543	0.4191	0.6966	0.5110
Verb (all)	0.7131	0.5568	0.6023	0.4001	0.5764	0.3631	0.5829	0.4159
Light Verb	0.4919	0.3256	0.4538	0.2339	0.4310	0.2477	0.4873	0.3139
Mid-Range Verb	0.7160	0.5766	0.5927	0.4140	0.6301	0.4175	0.6271	0.4797
Rare Verb	0.8955	0.6485	0.8195	0.4592	0.8571	0.5000	0.8750	0.4218

Table 6: Automated Translation: Average HHI scores taken for words in automated translations as compared with the scores from human translations. Scores are provided in the range of 0 to 1

still areas in which consistency is a real problem, but one needs to look more closely at the data to find the problems.

Any consistency in the output of an SMT system will be accidental, and not by design. It is a reflection of the data that the system was trained with and represents the "best" choice for translating a word or phrase, as determined by scores from the phrase table and language model. Carpuat and Simard (2012) suggest that consistency in the source side local context may be sufficient to constrain the phrase table and language model to produce consist translations. It is also important to note that the outcome is very much dependent on the system used to perform the translation. Carpuat and Simard (2012) suggest that weaker SMT systems (i.e. those that report lower BLEU scores) may be more consistent than their stronger counterparts due to fewer translation options.

There are several possibilities. A word in the source language may be translated:

- Completely consistently (HHI = 1);
- Very inconsistently (HHI $\sim$ 0);
- or anywhere in between

Additionally, a translation that is deemed to be completely consistent may be either correct or incorrect. With humans, we assume the translation output to be of a high standard but we cannot assume the same of an SMT system.

Examples of completely consistent translations are *horses* as "chevaux", *man* as "homme" and *nails* as "clous". All are taken from *Dracula*. While *horses* and *man* are translated correctly, "clous" is an incorrect translation of *nails* which the context of the novel refer to Dracula's fingernails. "ongles" would have been the correct translation. The word "clous" is typically used in the sense of nails used in construction. This is an example of a translation that could result either from lack of sufficient local context (for disambiguation) or because "ongles" is not present in the data the SMT system was trained on.

Examples of inconsistent translations are for the body parts *arm* and *hand* in the text of *Dracula*. *arm* is translated either correctly as "bras" (arm, body part) or incorrectly as "armer" (the verb "to arm"). *hand* is translated correctly as "main" ("hand") and incorrectly as côté ("side") and "part" ("portion"). In both cases, the correct translation was available to the system and a more accurate translation could have been obtained had the correct translation been identified and its consistency encouraged.

Ambiguous words in particular can cause trouble for SMT systems. There are many words that can function as both a verb and a noun, e.g. *process* and *count*. Local context might not always be sufficient to provide the correct disambiguation, resulting in opportunities for incorrect translations.

An example of where an ambiguous word results in problems is in the translation of *count* (i.e. Count Dracula) as: omitted (4), "compter" (21), "compatage" (2), "comte" (1) and "dépouillement" (5). The only acceptable translation from this set is "comte". As for the reaming options: "compte" and "comptage" are both verbs meaning "to count" and "dépouillement" is a noun meaning "starkness", "austerity" or "analysis" (of data).

5 Conclusion

The analysis of human translation presented in this paper is a first attempt to understand where and when it might be appropriate to encourage consistency in an SMT system. I consider genre as the *where* and parts-of-speech as the *when*, but other interpretations are also possible. On the whole, it seems reasonable to encourage the con-

sistent translation of nouns, across all genres. In addition, encouraging consistency in the translation of rare verbs and adjectives for technical documents and of adjectives for public information documents may also prove beneficial.

With respect to verbs, variation in verb consistency has been shown to correlate with frequency (as a proxy to identify light and rare verbs). Given the low consistency with which humans translate light verbs, encouraging their consistency in automated translation would be undesirable.

Automated translation may look very consistent on the surface, but it is necessary to look beyond this to see the errors. While humans may make inconsistent translations, we trust that these inconsistencies will not confuse or mislead the reader. SMT systems on the other hand generate their translations based on statistics that say what the "best choice" might be, both at the word/phrase level (through the phrase table) and overall (through the language model). Furthermore, they do nothing to guarantee consistency - this occurs by chance, whether desirable or not. As a result, inconsistencies may arise that make the translations difficult to read. These inconsistencies are not predictable and could occur in any SMT system.

6 Future Work

The findings presented in this paper are suggestive but only a small number of texts have been included for each genre. The analysis could be extended to include a larger set of documents and different language pairs (the only requirement is for a POS tagger for the source language). Multiple translations of the same document could also be considered to identify whether similar patterns can be observed for different translators.

There are a number of possible ways in which to use this information to inform the design of a SMT system. I have shown that SMT systems are capable of highly consistent translations but this consistency cannot be guaranteed and there is the possibility that the translations will be consistent and incorrect. Also, Carpuat and Simard (2012) have shown that inconsistent translations in SMT often indicate translation errors. A system which encourages translations which are both consistent and correct (or at least acceptable) for words that belong to a predefined set (e.g. by POS tag) is desirable. This "encouragement" could be achieved using rewards delivered via feature functions or within n-best list re-ranking – hypotheses which make re-use of the same translation(s) for repetitions of the same source word would be ranked higher than those that introduced inconsistencies. Revisiting the cache-based models of (2010a; 2010b) and Gong et al. (2011) could provide a possible starting point.

The initial focus could be on nouns, which are translated by human translators with high consistency for all genres. Many nouns are used either to specify entities that are only mentioned once in a text (essentially setting the scene for more prominent entities), or as "predicate nominals" on those more prominent entities (e.g. in "...is a horrific *story*"). However, other nouns occur within the Noun Phrases (NPs) that make up part of a *coreference chain*, of subsequent reference to prominent entities.

As an extension to this work I will aim to investigate the consistency of translation of those nouns that belong to coreference chains and ultimately, to build a system that makes use of the resulting information. Work has already started to construct a parallel corpus in which coreference chains are annotated so that the translation of coreference (both NPs and pronouns) may be studied in more depth.

Another question worth considering is whether it would be desirable to replicate aspects of low consistency in human translation by encouraging inconsistent (but still acceptable) translations of certain words or word categories. My instinct is that this could lead to translations that better approximate those produced by humans.

7 Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE). Thanks to Professor Bonnie Webber for her guidance and numerous helpful suggestions and to the three anonymous reviewers for their feedback.

References

Anne Abeillé, Lionel Clément, and Alexandra Kinyon. 2000. Building a treebank for french. In *In Proceedings of the LREC 2000*.

Marine Carpuat and Michel Simard. 2012. The trouble with smt consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT

'12, pages 442–449, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeremy H. Clear. 1993. The digital word. chapter The British national corpus, pages 163–187. MIT Press, Cambridge, MA, USA.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, pages 233–237, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 909–919, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Yifan He, Yanjun Ma, Andy Way, and Josef van Genabith. 2011. Rich linguistic features for translation memory-inspired consistent translatio. In *Proceedings of Machine Translation Summit XIII*, pages 456–463.

Masaki Itagaki, Takako Aikawa, and Xiaodon He. 2007. Automatic validation of terminology consistency with statistical method. In *Proceedings of Machine Translation Summit XI*, pages 269–274. European Associaton for Machine Translation.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT, AAMT.

Kazem Lotfipour-Saedi. 1997. Lexical cohesion and translation equivalence. *Meta: Journal des Traducteurs / Meta: Translators' Journal*, 42(1):185–192.

Yanjun Ma, Yifan He, Andy Way, and Josef van Genabith. 2011. Consistent translation using discriminative learning: a translation memory-inspired approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1239–1248, Stroudsburg, PA, USA. Association for Computational Linguistics.

I. Dan Melamed. 1997. Measuring semantic entropy. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics*, pages 41–46.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48, March.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51, March.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. In *IN ARXIV:1104.2086*.

Raphael Salkie. 2010. The intersect translation corpus. Available on the web: http://arts.brighton.ac.uk/staff/raf-salkie/portfolio-of-major-works/intersect.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Jörg Tiedemann. 2010a. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 8–15, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jörg Tiedemann. 2010b. To cache or not to cache? experiments with adaptive models in statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 189–194, Uppsala, Sweden, July. Association for Computational Linguistics.

Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 417–426, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tong Xiao, Jingbo Zhu, and Shujie Yao. 2011. Document-level consistency verification in machine translation. In *Proceedings of MT summit XIII*, pages 131–138.