Embedding register-aware MT into the CAT workflow

Corey Miller
The MITRE Corporation, McLean, VA 22102 camiller@mitre.org
Danielle Silverman danielle.c.silverman@nvtc.gov
National Virtual Translation Center, Washington, DC 20535
Vanesa Jurica vjurica@mitre.org
The MITRE Corporation, McLean, VA 22102
Elizabeth Richerson liz@mitre.org
The MITRE Corporation, McLean, VA 22102
Rodney Morris rodney.d.morris@nvtc.gov
National Virtual Translation Center, Washington, DC 20535
Elisabeth Mallard Elisabeth.d.mallard@nvtc.gov
National Virtual Translation Center, Washington, DC 20535

Abstract

As machine translation (MT) improves, the possibility for it to translate different registers appropriately becomes more possible. This capability is particularly relevant when confronting non-standard varieties such as are common in social media and chats. Register-sensitive MT, coupled with advances in register detection, opens up new possibilities for the enhancement of computer-assisted translation (CAT) tools.

1. Introduction

Many translation style guides say something like the following: “A translation is not just a transcription from one language into another. It needs to render not only the meaning of words and sentences but also the context and, more subtly, what is sometimes described in stylistic manuals as the register of the source text—its level and style of language.” (World Bank 2004). It should be noted that such advice is also proffered to interpreters: “From the standpoint of the user, a successful interpretation is one that faithfully and accurately conveys the meaning of the source language orally, reflecting the style, register, and cultural context of the source message, without omissions, additions or embellishments on the part of the interpreter” (Federal Coordination and Compliance Section, 2011).

If such advice is to be adhered to, it will of course be most challenging to those translation departments whose work spans a wide range of registers, from slangy/chatty to scientific/formal. After discussing the nature of register and the motivation for advising its conveyance from source to target, we seek to establish ways in which computer-assisted translation (CAT) software can aid translators in this process. In particular, we explore the role of both termbases and translation memories (TMs) in recording register information. Within that context, we consider the possible role of automatic register detection. Finally, we discuss work on register in machine translation (MT) and how this may ultimately facilitate a register-enabled CAT workflow.

2. Register

Ability to effectively translate register variation is relevant to all spheres of translation, whether governmental, corporate or literary/creative. Government organizations must translate from a variety of sources, possibly ranging from social media to diplomatic communications and businesses produce communications variously targeted for customers and partners. Perhaps register variation is most obvious in the creative domains, including subtitling and fiction. In all of these domains, even literary ones (Francisco 2015), CAT usage is becoming more prevalent.

Failure to convey register in translation can have real-life consequences, even in "translations" within a given language. Jackman (2017) describes the case of Warren Demesme, who said "just give me a lawyer dog" in the course of his police interrogation in Louisiana. The Louisiana Supreme Court decided that this utterance did not constitute an invocation of the right to counsel. As pointed out by Green (2002), *dog* or *dawg* is a term of address used by African American males, "without negative import", a detail neither the interrogators nor the court seems to have taken into account.

Steiner's (1998) register-based analysis of original English and subsequent German Rolex advertisements indicates that adherence to register across languages may have business consequences. From this analysis, it appears that the effectiveness of the advertisements may be compromised by a failure to convey source language register to the target language in certain cases.

2.1. Definition

While at first glance, register might seem to simply include stylistic variation, we feel it may be useful to explore whether what is intended in recommendations to convey register from source to target might really include the full range of sociolinguistic variation. In addition to style, sociolinguistic variation can include language differences associated with geography (from nation to neighborhood), age (from youth to old), as well as social, ethnic or religious affiliation.

The notion of markedness proves useful in this discussion. While the notion of "standard" language is somewhat fraught, it seems fairly straightforward to say that a given expression is marked for a particular sociolinguistic category. For example, while a word like "money" might be considered neutral in most, if not all, varieties of English, a word like "dough" in the sense of "money" seems marked for informality. In the same way, "elevator" is markedly American (and perhaps beyond) and "lift" is markedly British (and perhaps beyond).

First, it will be helpful to assess whether register covers all the sociolinguistically-relevant information we might want to convey. Halliday (2002) focuses on functional (diatypic) registers, i.e. the language differences encountered across domains. Yu (2017) has a wider conception, including the range of language from vulgar to elevated, and the possibility both of several registers in a given domain, and the preponderance of certain registers within certain social groups.

Dialect or locale introduces another axis of sociolinguistic variation that is hard to dissociate from register, and its conveyance from source to target is not straightforward. For example, while "pop" and "soda" are two locale-specific ways of referring to carbonated beverages in American English, how or should one convey this in a French translation? Hanes (2012) discusses the rendering of Southern American English into Brazilian Portuguese subtitles and finds a variety of available strategies, many of which could benefit from a more structured approach to this problem.

Finally, we consider the sociolinguistic approach to style variation. Labov (2006) offers the notion that style is correlated with attention paid to speech, and work in this tradition explores a range of styles from casual to formal based on the task at hand, e.g. conversation vs. reading word lists. Bell (1984) offers a slightly different perspective, viewing style as “audience design”, thus shifting the focus from attention to accommodation to interlocutors. We find that both perspectives on style variation are useful here.

For our purposes, the attentional view of style provides insight into the kinds of deviations from neutrality that often perplex translators: typographical errors, disfluencies, malaprops, and in the case of non-native speakers/authors, false friends (Chamizo-Domínguez 2008). At the same time, the audience-focused view of style is relevant when we consider for whom the translation is intended.

While we will not go into further detail on this topic here, we feel a proper analysis of how such displays of lack of attention (or education/experience/training) should be conveyed in translation is a proper part of register analysis and conveyance. While treatment of such “errors” are often handled by a diverse set of tools from use of the term *sic* to subtle correction in target language translations, it is clear that each of these devices carries baggage, whether by impugning the source author/speaker or masking/“upgrading” the source author’s intent and characteristics.

3. Terminology Management

Termbases are a natural repository for register information. ISO/TC37/SC3, “Systems to manage terminology, knowledge and content” includes ISO/CD 12620, “Terminology and other language and content resources -- Data category specifications” whose current data categories are described at http://www.datcatinfo.net. The register, dating and frequency data categories shown in Table 1 provide a starting point for considering the extent to which dialect and register issues can be dealt with through terminology management.

Dating	Modern
	Old
Frequency	Commonly used
	Infrequently used
	Rarely used
Register	Bench-level
	Dialect
	Facetious
	Formal
	In house
	Ironic
	Neutral
	Slang
	Taboo
	Technical
	Vulgar

Table 1. Register data categories in ISO/CD 12620

While fairly extensive, this list is not exhaustive. For example, age grading, i.e. the use of terms by certain age groups and not others, sociolects and dialect/locale need to be (more finely) addressed. So, in contrast to simple data categories, these could allow more complex data types.

4. Translation Memory

TM metadata is another potential repository for register information, but a survey of the relevant standards (e.g. TMX, XLIFF) and literature indicates that it is not common to characterize translation units (TUs) beyond their date, translator, status and domain. Moorkens (2013) indicates the usefulness of date metadata, since more recent TUs may benefit from having been corrected and feature the latest vocabulary. However, even if register information were made specifiable in TMs, analogous to that proposed in the standards for terminology, it remains to be established how and whether this could be made useful for translators in conveying register into the target language.

In our imagined scenario, source and target variants of translation units in a translation memory would have the same metadata elements available to them, but they would not necessarily match between source and target. For example, while elements such as “frequency” or “slang” might well match, if further sociolinguistic detail is provided, such as “Southern United States” for the American English variant of a TU, this cannot be expected to have the same value as the equivalent in another language.

At this stage, we are agnostic as to whether separate TMs should be maintained for register. Given the potentially large dimensionality of register information, it seems that such an approach would prove cumbersome. At any rate, it seems that a register-sensitive TM search should not be greatly affected by the choice to use single or multiple TMs.

4.1 Automatic Register Detection

Assuming that register information can be captured in a TM, the question arises as to what extent this can be done automatically as new TUs are added. Lapshinova-Koltunski & Vela (2015) and Biber (2014) discuss automatic identification of registers. Salloum et al. (2014) discuss sentence-level dialect identification in service of MT; this seems like a promising approach to apply in the case of identifying register in source materials, since it can shift multiple times over the course of the document. Once enhanced with register information, TUs could be compared for register across languages, resulting in a way for translators and reviewers to identify register mismatches warranting further attention. Register detection could also enhance content optimization tools’ capabilities with respect to style checking.

5. Machine Translation

Since MT can help fill in the gaps in TM, it is worth exploring whether it too can be made sensitive to register. Niu et al. (2017) describe initial attempts at imbuing MT with the ability to control register. In cases without a suitable TM match, MT register control could be exercised on source language TU variants in order to produce target variants with the appropriate register.

6. Discussion

How might all of this work in practice? Let us imagine a workflow including a CAT tool of the future which we call Register CAT. Register CAT provides both terminology management and translation memory both enhanced with register metadata, and the translation memory is enhanced with register detection capabilities. The translation memory backs off to machine translation that can output various registers on demand.

At this stage, we imagine that an input source text for Register CAT is segmented into sentence-sized segments as is common in today's CAT tools. While it is clear that register can vary within documents, it is not clear whether a different kind of segmentation would serve register conveyance better, and indeed, as will be seen below, individual segments can exhibit register variation as well.

The source language part of the translation unit can then be submitted to a language-specific register detector which can fill in register metadata information. This can then be used to search the translation memory, which has also been annotated (either by hand or by a register detector) for register. We leave the theory and mechanics of the factoring in of register information to TM match scoring to further research, but for now, we assume that better register matches will score higher than otherwise equally matching target language variants.

As the translator sets about modifying the translation memory match, the termbase will provide guidance by presenting register-matched target terms highest in the list. Once the translator is ready to commit the translation unit, the opportunity will be made available for her to modify the translation unit metadata in case the automatically generated register information requires it. In those cases where there is no suitable translation memory match for the source language segment, register-enabled MT will provide a target language segment best matching the source language register specifications.

Let us work through an example based on an American English news report from Oklahoma (https://www.youtube.com/watch?v=ydmPh4MXT3g). For the sake of this example, we will assume that the audio has been transcribed into the source language orthography and that the transcript is what constitutes the source side presented in Register CAT that needs to be translated into a target language, e.g. French.

> (Announcer) One resident describes her horrifying experience when she first realized the complex was on fire.
>
> (Sweet Brown) Well, I woke up to go get me a cold pop…

The announcer's utterance displays no marked properties and could well be described as register-neutral. In contrast, Sweet Brown's utterance displays two features that can be considered marked with respect to American English: "go get me" and "pop". The use of "me" in the phrase "go get me" is an example of a personal dative which Horn (2008) ascribes to "dialectal (Southern and Appalachian) U.S. English". The use of "pop" vs. "soda" (among other variants) has been a longstanding discussion among American English dialectologists (von Schneidemesser 1996). According to popvssoda.com, an internet survey project by Alan McConchie, "pop" is the lead variant in Oklahoma County, where the broadcast emanates. In fact, the state of Oklahoma appears to be one of the southernmost regions for "pop", whose main bastions appear to be the Northwestern and North-central regions of the United States.

Whereas the personal dative in "go get me" has a vernacular flavor, "pop" seems to be more of a geographical rather than a stylistic variant. In termbase metadata, we could indicate the stylistic and geographic features of each expression:

> "go get me": informal, dialect:Southern/Applachian (US)
> "pop": neutral, dialect:Northwestern/North-central (US), including Oklahoma

However, when considering how to indicate the metadata in the TM for the entire segment "Well, I woke up to go get me a cold pop…", we propose to take the union of the register features exhibited by the expressions it contains. Indeed, one segment could certainly contain both informal and formal words, or words with different geographical affiliations. Therefore, it is important to allow the metadata to accommodate this, perhaps by quantifying the number of expressions in a segment containing each relevant feature. This creates a situation where some utterances will have stronger sociolinguistic marking than others. For the expression in question, this could look something like this:

> "Well, I woke up to go get me a cold pop": informal (1), Southern/Appalachian US (1), Northwestern/North-central/Oklahoma (1)

Now when considering a translation memory search for a target language, say French, equivalent for this phrase, we confront all of the sociolinguistic variation of that language. In this case, the personal dative, as in "Je me prends un petit café", literally "I take me a little coffee" (Horn 2008) does not seem to be markedly informal as in American English. In the case of "pop", French has a number of terms of its own, as shown in Table 2.

boisson gazeuse	Formal
liqueur[1]	Canadian, Informal?
soda	Informal?

Table 2. French equivalents for "pop/soda"

While mapping register properties like "formal" and "informal" between languages may seem at first straightforward, we are confronted by the lack of register parallelism (at least with respect to vocabulary) between languages. So, if "go get me" is informal in English, and the equivalent with a personal dative is not marked in French, must we try to find an informal way of expressing that? If "pop" is regionally marked in English, is our translation best served by seeking a regionally-marked term like "liqueur" in French? We do not offer solutions to these problems here, but they are discussed elsewhere (e.g. Berezowski 1997).

It is hoped that this outline specifying ways in which the CAT workflow can be fortified to accommodate register information will provide researchers and developers a path forward for a forthcoming generation of CAT tools which will make it easier for translators and reviewers to maintain and assess register fidelity.

References

Bell, Allan. 1984. Language style as audience design. Language in Society 13, 145-204.

Berezowski, Leszek. 1996. Dialect in Translation. Wydawnictwo Uniwerytetu Wroclawskiego.

[1] Thanks to Miguel Jetté for Canadian French consultation.

Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. Languages in Contrast 14:1, 7-34.

Chamizo-Domínguez, Pedro J. 2008. Semantics and pragmatics of false friends. Routledge.

Federal Coordination and Compliance Section. 2011. Commonly Asked Questions and Answers Regarding Lim-ited English Proficient (LEP) Individuals. https://www.lep.gov/faqs/042511_Q&A_LEP_General.pdf [last accessed 25 September, 2017].

Francisco, Reginaldo. 2015. CAT tools em tradução literária: para quê? VI Congresso Internacional de Tradção e Interpretação da ABRATES.

Green, Lisa J. 2002. African American English: A Linguistic Introduction. Cambridge University Press.

Halliday, M.A.K. 2002. The construction of knowledge and value in the grammar of scientific discourse: With reference to Charles Darwin's The Origin of Species. In Jonathan J. Webster, editor, Linguistic Studies of Text and Discourse, Volume 2, Bloomsbury, pages 169-192.

Hanes, Vanessa Lopes Lourenço. 2012. Norms in the Translation of Southern American English in Subtitles in Brazil: How is southern American speech presented to Brazilians? Translation Journal 16(3). http://translationjournal.net/journal/61southern.htm [last accessed 25 September, 2017].

Horn, Laurence R. 2008. "I love me some him": The landscape of non-argument datives. In O. Bonami and P. Cabredo Hofherr, editors, Empirical Issues in Syntax and Semantics 7, pages 169-192.

Jackman, Tom. 2017. "The suspect told police 'give me a lawyer dog.' The court says he wasn't asking for a lawyer." Washington Post, November 2. https://www.washingtonpost.com/news/true-crime/wp/2017/11/02/the-suspect-told-police-give-me-a-lawyer-dog-the-court-says-he-wasnt-asking-for-a-lawyer.

Labov, William. 2006. The social stratification of English in New York City, Second Edition. Cambridge.

Lapshinova-Koltunski, Ekaterina and Mihaela Vela. 2015. Measuring 'registerness'in human and machine translation: A text classification approach. Proceedings of the Second Workshop on Discourse in Machine Translation, pages 122-131.

Moorkens, Joss. 2013. The role of metadata in translation memories. In Valerie Pellatt, editor, Text, Extratext, Metatext and Paratext in Translation, Cambridge Scholars Publishing, pages 79-90.

Niu, Xing, Marianna Martindale and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2804-2809.

Salloum, Wael, Heba Elfardy, Linda Alamir-Sallou, Nizar Habash and Mona Diab. 2014. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 772-778.

Steiner, Erich. 1998. A register-baed translation evaluation: An advertisement as a case in point. Target 10:2, 291-318.

von Schneidemesser, Luanne. 1996. Soda or Pop? Journal of English Linguistics, 24(4): 270-287.

World Bank. 2004. World Bank Translation Style Guide, Version 1.0. http://siteresources.worldbank.org/TRANSLATIONSERVICESEXT/Resources/Translation_Style_Guide_English.pdf [last accessed 25 September, 2017].

Yu, Jing. Translating ‘others’ as ‘us’ in Huckleberry Finn: dialect, register and the heterogeneity of standard language. Language and Literature, 26(1): 54-65.