Routledge
Taylor & Francis Group

International Interactions

Empirical and Theoretical Research in International Relations

ISSN: 0305-0629 (Print) 1547-7444 (Online) Journal homepage: www.tandfonline.com/journals/gini20

Translating CAMEO verbs for automated coding of event data

Javier Osorio, Viveca Pavon, Sayeed Salam, Jennifer Holmes, Patrick T. Brandt & Latifur Khan

To cite this article: Javier Osorio, Viveca Pavon, Sayeed Salam, Jennifer Holmes, Patrick T. Brandt & Latifur Khan (2019) Translating CAMEO verbs for automated coding of event data, International Interactions, 45:6, 1049-1064, DOI: 10.1080/03050629.2019.1632304

To link to this article: https://doi.org/10.1080/03050629.2019.1632304

INTERNATIONAL INTERACTIONS
2019, VOL. 45, NO. 6, 1049–1064
https://doi.org/10.1080/03050629.2019.1632304

Routledge
Taylor & Francis Group

NOTE

Translating CAMEO verbs for automated coding of event data

Javier Osorio [a], Viveca Pavon[a], Sayeed Salam[b], Jennifer Holmes[b], Patrick T. Brandt[b], and Latifur Khan[b]

[a]University of Arizona; [b]University of Texas

ABSTRACT

The Conflict and Mediation Event Observations (CAMEO) framework is the industry standard for computerized event coding that produces massive data on political conflict at a global scale. However, despite their focus on conflict in foreign locations, most coding efforts have the limitation of exclusively relying on text written in English. This paper advances the frontier of computerized event coding by presenting the CAMEO Verb Translation Application (VTA), a new technology to enable the translation of CAMEO verb dictionaries into non-English languages. CAMEO VTA integrates automated translation with human-in-the-loop features to increase the validity of the translation task. This development opens the possibility of expanding the leverage of CAMEO in a systematic and efficient manner. The first application of this tool focuses on translating CAMEO into Spanish.

KEYWORDS

Event data; CAMEO; Spanish translation; human in the loop

International Relations and Political Science have benefited from recent technological innovations and the availability of big data to study conflict processes around the globe. Although there are reasons for enthusiasm about the opportunities that these innovations open (de Vries, Schoonvelde, and Schumaher 2018; Ulfedler 2013), we advance a cautionary note when using automated translation tools to generate event data from non-English languages. By focusing on Spanish, this paper presents new technology for translating the Conflict and Mediation Event Observations (CAMEO) framework (Gerner, Schrodt, and Ömür 2009), which is the industry standard for computerized event coding. While taking advantage of automated translation tools, this technology incorporates human judgment to reduce automation-induced errors. As this research shows, automated translation is powerful for expanding text-as-data analysis to non-English languages, but the input from

CONTACT Javier Osorio josorio1@email.arizona.edu University of Arizona

This research was financially supported by the National Science Foundation [Award 1539302]. We thank the reviewers and colleagues for their invaluable contributions to the research. We thank our research assistants for their superb support: Erick Alonzo, Yamilette Peguero, Normary Alcivar, Yaritma Cabral, Jonathan Panuela, Diego Morales, Danielle Edmonds, Luciana Arteaga, Andrew Almaguer, Sebastian Poveda, Manuel Garcia, Maryam Bahojb, and Mansi Savalia.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/gini.

human experts still is crucial for improving coding validity, particularly for domain-specific research such as conflict studies. In this way, combining automated translation with human-in-the-loop (HITL) allows advancing more confidently toward multilingual event coding to better study conflict processes in foreign locations.

For the last 20 years, CAMEO has grown and consolidated as the main framework for event data coding. CAMEO's early applications though the TABARI coder (Schrodt 2009) launched a prolific conflict research agenda using event data (Schrodt 2001, 2006, 2012; Schrodt and David 2013; Schrodt, Davis, and Weddle 1994; Shellman, Hatfield, and Mills 2010). By covering a broad range of cooperative or conflictive behaviors of verbal or material nature, CAMEO became cornerstone of the most relevant event data projects to date: the Integrated Conflict Early Warning System (ICEWS) (O'Brien 2010) and the Open Event Data Alliance's (OEDA) Phoenix data using PETRARCH (Schrodt, Beieler, and Idris 2014).

However, these coding schemes are limited by their dependency on English language. For example, Phoenix only processes English writing and ICEWS processes foreign languages by using automatic translation into English before coding the events, which could be problematic. Surprisingly, despite these projects' ambition to code conflicts occurring in foreign locations, they do not directly rely on information written in foreign languages. The problem of exclusively relying on English sources to code foreign location events is the risk of incorporating coverage bias from news sources, thus reducing the quality of the event data derived from these stories.

Developing multilingual coding technologies provide several contributions to tracking conflict processes in foreign locations. Such developments facilitate gaining direct insights from local sources. They also eliminate the need for translation or using analysts with foreign language skills. Multilingual event coding also reduces concerns of bias from English language sources, thus improving the inferences drawn from international conflict data sets and refining the precision of conflict forecasting (Brandt, Freeman, and Schrodt 2011; Herkenrath and Knoll 2011). Recent Natural Language Processing (NLP) developments conducted by the Open Event Data Alliance[1] are advancing toward multilingual event coding, thus facilitating the analysis of international conflict processes. These developments contribute to other multilanguage NLP technologies such as Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Brthard, and David 2014) and Eventus ID (Osorio and Reyes 2016).

This paper pushes the technological frontier one step forward by presenting the CAMEO Verb Translation Application (VTA) — generic web-based platform to assist the translation of CAMEO verbs dictionary. This

[1]See http://github.com/openeventdata (accessed on 4/19/2019).

technology allows expanding CAMEO's potential into non-English languages without developing a coding dictionary from scratch and while keeping the comparability of coding categories across languages. CAMEO VTA includes HITL features to enable human validation for assessing the accuracy of automated translation in the context of CAMEO categories. As the application shows, despite the contributions of computer-assisted translation, human coders still play a crucial role in guaranteeing the validity and accuracy of the translated dictionaries. The particular application of this paper focuses on documenting the translation of CAMEO into Spanish.

Future endeavors can expand CAMEO to other non-English languages with this infrastructure. Ultimately, the researchers of the Open Event Data Alliance will use the Universal PETRARCH coding engine (a successor to Schrodt's earlier TABARI and PETRARCH coders), which relies on universal dependencies to code event data using the CAMEO event ontology. This application will be deployed to code events from English, Spanish, and Arabic text.

CAMEO Verb Translation Application

CAMEO VTA was developed by computer scientists at the University of Texas – Dallas (UTD) with inputs from Spanish language teams at John Jay (JJ) College and UTD. This tool is a web-based interface for collaborative translation of CAMEO verb dictionaries. The system takes advantage of WordNet (Fellbaum 1998; Miller 1995), a prominent online linguistics tool, to assist in a systematic translation of verb phrases comprised in the CAMEO dictionary. CAMEO VTA uses WordNet in English to identify verb definitions and examples and uses Wordnet in Spanish to suggest translated words to human coders for their validation. Enabling HITL features makes this system collaborative rather than fully automated. The following subsections provide an overview of CAMEO VTA's architecture and explain the use of its interface.

CAMEO VTA System

The CAMEO VTA system comprises a sequence of automated steps operating in the back-end, yet it requires HITL input in the front-end. Figure 1 presents these main steps, followed by a succinct explanation of the tasks conducted in each of them.

- **Step 1**. CAMEO VTA loads the CAMEO verb dictionary in English and uses it as the prime input file. The application then takes each related word from the dictionary.
- **Step 2**. For each word, the application searchers into WordNet related linguistic characteristics to identify sets of cognitive synonyms (known

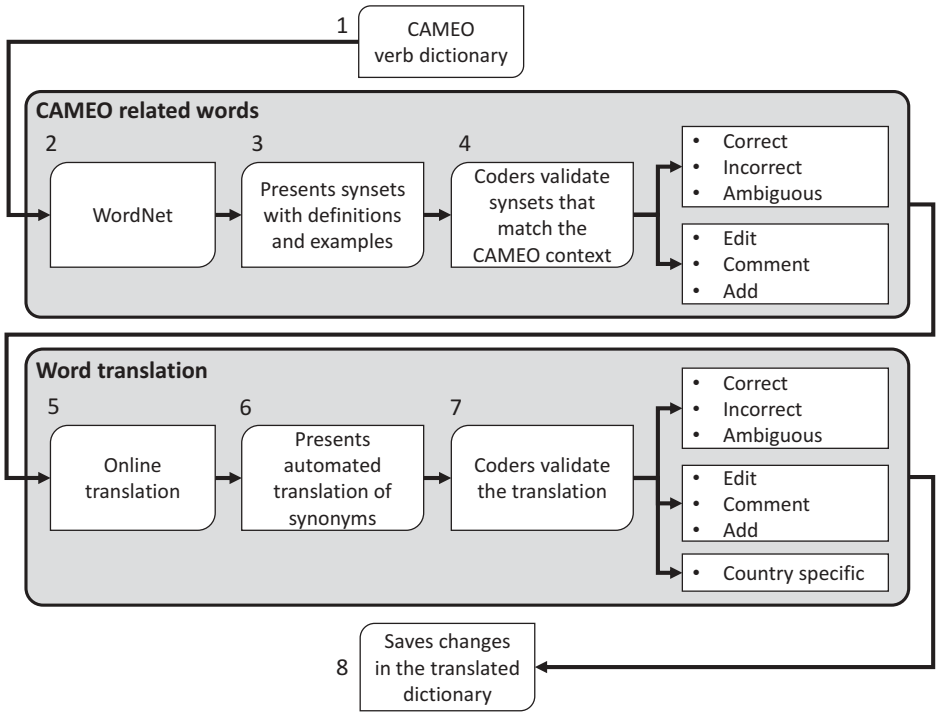


Figure 1. CAMEO VTA system.

as *synsets*), that express conceptual, semantic and lexical relations to the selected word.

- **Step 3**. Using the WordNet results, the interface presents to the coders a set of English definitions and examples that correspond to each word.
- **Step 4**. The HITL feature allows humans to validate the relevance of the WordNet definition and examples that match the overall concept of the CAMEO category. Users issue a verdict (correct, incorrect, or ambiguous) about the proposed definition and example of the synset, and are able to modify the synset content to better fit the underlying concept of the CAMEO dictionary.
- **Step 5**. The VTA system takes the CAMEO related word in English and looks for its corresponding Spanish translation using WordNet in Spanish. Since lexicons are complex and often nuanced, the system also considers any synonyms of the word translated into Spanish.
- **Step 6**. The system then presents the set of Spanish-translated verbs in the interface displaying the verb is infinitive form. This verb form simplifies translating conjugations from English into Spanish.
- **Step 7**. The HITL feature then allows coders to validate the accuracy of the Spanish translation by issuing a verdict (correct, incorrect, or ambiguous) with respect to the CAMEO category. Coders can also edit or make notes about proposed translated words or create new ones. The

interface also allows identifying colloquialisms to indicate country-specific use of particular words.

- **Step 8**. Finally, the system saves the changes in the Spanish dictionary.

CAMEO VTA automatically implements steps 1, 2, 3, 5, and 6 and the HITL feature loads this information in the interface so humans provide their verdict in steps 4 and 7, and save the changes in step 8.

CAMEO VTA Interface

The CAMEO VTA system deploys the interface online, and it is designed for simultaneous collaborative work of multiple users from remote locations.[2] Also, having multiple coders allows various rounds of revisions of each CAMEO related word that can be used to assess intercoder reliability for the translation tasks.

Coders start the session by logging into the system with a unique ID. Then users select a CAMEO category to work with from the CAMEO List on the web interface. The CAMEO List displays the 20 different CAMEO verb categories,[3] which can be disaggregated by their specific sub-codes. For example, Figure 2 depicts the two-digit categories for codes 17: Coerce, 18: Assault, 19: Fight, and 20: Use Unconventional Mass Violence. The menu also displays the three-digit sub-categories for code 19: Fight. Code 195 has the radial button closed, but it can be further expanded to display the four-digit CAMEO sub-categories.

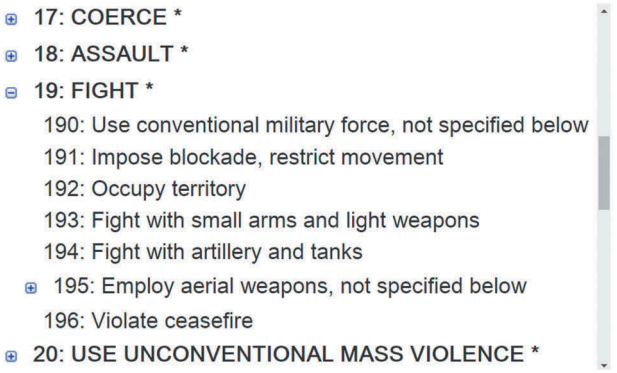


Figure 2. CAMEO VTA list.

[2]The CAMEO VTA application is available online in the Open Event Data Alliance Github site (https://github.com/openeventdata/synset_validator) once the manuscript is published.

[3]CAMEO includes the following action categories: 01: make public statement, 02: appeal, 03: express interest to cooperate, 04: consult, 05: engage in diplomatic cooperation, 06: engage in material cooperation, 07: provide aid, 08: yield, 09: investigate, 10: demand, 11: disapprove, 12: reject, 13: threaten, 14: protest, 15: exhibit force posture, 16: reduce relations, 17: coerce, 18: assault, 19: fight, 20: use unconventional mass violence.

Once users select a specific code, they can explore its details to gain more familiarity with its concept as outlined in the corresponding CAMEO codebook entry in English. The codebook includes information such as the name of the code, description, usage notes, and coding examples.

They way the CAMEO dictionary is structured, each code category includes a set of verbs that capture different aspects of the underlying concept of the code. These verbs are known in the coding protocol as *related words*, as they relate to the overall meaning of the corresponding CAMEO code. After gaining familiarity with the code details in the previous step, users choose a related word from the drop-down menu. This procedure loads the interface with all the different synsets in English gathered from WordNet, along with their meanings and examples. This process corresponds to steps 1–3 indicated in Figure 1. For each verb in the English synset, the VTA also loads a set of translated words extracted automatically from WordNet in Spanish. This process corresponds to steps 5 and 6 as depicted in Figure 1. Each English verb might have multiple correspondences in Spanish as WordNet identifies synonyms in a network of meanings. For example, Figure 3 is the coder's partial view of the interface for code 190: Fight using the related word "Aim".

The first HITL coding task is to evaluate the validity of each specific English synset with respect to the overall meaning of the CAMEO code as indicated in the codebook details window. In this stage, users can offer feedback, edit, or add content to the definition and examples of each synset, and provide a verdict (correct, incorrect, or ambiguous) about its suitability with respect to the overall meaning of the CAMEO code. This process corresponds to step 4 indicated in Figure 1.

The second HITL task is to evaluate the validity of the Spanish translation of each English synset with respect to the meaning of the CAMEO code. To do so, users mark each Spanish entry as correct, incorrect, or ambiguous. Coders are also able to provide country-specific usage and add new words in Spanish if necessary. These features are particularly important for including colloquialism and slang words commonly used in news reports but that

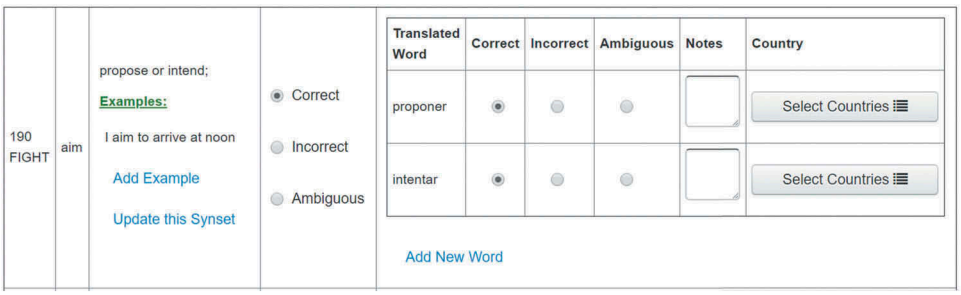


Figure 3. CAMEO VTA Interface.

rarely appear in formal language dictionaries such as WordNet or *Real Academia Española*. This process corresponds to step 7 as shown in Figure 1.

Training Coders

CAMEO VTA includes features that provide automated suggestions for users to translate the verbs dictionary in a systematic manner. However, given the state of automated translation technology, the precision of the automated assistance is limited, and the accuracy of the translation process ultimately depends on human coder's language skills. In this application, the researchers relied on two teams of undergraduate students: five coders at John Jay College and six coders at the University of Texas at Dallas. The teams worked jointly in translating the CAMEO verb dictionary v2.0 from English to Spanish.

Coders were first screened for Spanish proficiency. To do so, researchers tested prospective coders to assess their language competences and see if they could accurately identify basic components of event data (the source, action, and target) in Spanish text. While many students are conversationally competent, some may not have sufficiently strong reading or writing skills to be able to conduct the research in the written realm — thus the need for a Spanish qualification test. The selected students vary in fields of study, country of origin, and gender, but all of them were proficient in Spanish.

Once team members are selected, they need to become familiar with the core components of event coding. Coders went through rigorous training that included understanding the event data structure and gaining familiarity with the CAMEO codebook and the role of CAMEO in UD PETRARCH. To do so, each student read 200+ stories in Spanish to identify the source, action, and target triad. The researchers and the students held meetings to discuss the results for each story. Then, selected coders received training in the CAMEO VTA system to evaluate the English synsets and Spanish translations for each CAMEO category.[4] Coders from each team conducted multiple rounds of translation of all CAMEO categories. In addition, researchers divided the CAMEO codes so one team would validate a set of CAMEO categories first, then switch with the task to the other team. This redundancy and coding swap is a valuable approach to generate intercoder reliability and identify the extent of agreement in each aspect of the translation protocol. It also became an important feature for validating new Spanish words that were being added by the first round of coders.

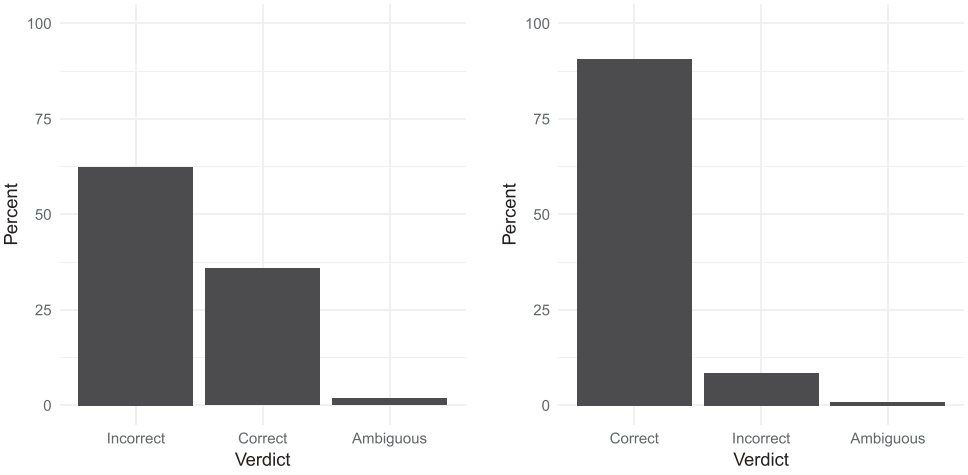
Translating CAMEO into Spanish

This section presents the main results of translating the CAMEO verb dictionary into Spanish. In total, coders issued 28,932 verdicts for the

[4]The Online Appendix includes the codebook used in the training process.

match between English synsets and the CAMEO categories (this feedback corresponds to Step 4 of the VTA process shown in Figure 1). As Panel (a) in Figure 4 indicates, coders assigned a correct verdict in only 35.8% of the English synsets (10,358 instances) gathered through WordNet. In contrast, the synset content did not align with the meaning of the CAMEO category in 62.5% of the cases, thus receiving a verdict of incorrect in 18,070 instances. According to the coders, about two-thirds of the verb definitions or examples of the English synsets extracted from WordNet did not apply to the context or the underlying concept of the CAMEO category and received the verdict of incorrect. This is about twice the proportion of synsets identified as correct. This result reveals the importance of having HITL to verify the validity of automated translation. This is particularly important for domain-specific coding projects in which expert validation is crucial for disambiguation. The centrality of HITL also serves as a cautionary note for the optimism toward automated translation (see de Vries et al. 2018), to the extent that the validity of the automated output largely depends on the input text. If such input text has not been curated by human experts, the validity of the automated translation is likely to contain substantial error, thus affecting the description and inferences made from the text-as-data.

Given that a single verb can have multiple definitions, it is expected that only a few of those definitions will match the CAMEO context and be categorized as correct synsets. For instance, in the CAMEO codebook, code 192 refers to the concept "Fight: occupy or seize control of a territory using armed forces." The first related word to translate is "capture" and one of its definitions is "bring about the capture of an elementary particle or celestial body and causing it enter a new orbit". This specific definition does not



(a) Verdict of English synsets

(b) Verdict of Spanish verb translation

Figure 4. Coder verdicts for synsets and translations.

match the CAMEO concept even though the abstract meaning of the word "capture" would. In this case, the coder would issue a verdict indicating that this definition or usage of "capture" is incorrect for the CAMEO code 192; then, the remaining translated words would not be taken into consideration in the translation scheme. Similarly, there are English words that, when translated into Spanish, have a definition that would be incorrect under the CAMEO code. For example, "to assault" in English, would be directly translated as "*asaltar*" in Spanish, but this word actually means "to rob."

The performance metrics also provide strong evidence for the advantages of involving HITL for collaborative translation efforts rather than relying on fully automated approaches. Given the low agreement of English verbs, definitions and examples extracted from WordNet (62.5%), it is clear that machine translations alone will introduce substantial error into the event data, if not corrected by humans.

After assessing the English synsets, coders validated the Spanish translation in 37,440 instances. As Panel (b) in Figure 4 shows, 90.1% of the Spanish verb translations were correct.[5] This metric indicates a high degree of reliability of automated Spanish translation of the verb in infinitive, contingent on humans selecting an English verb that matches the CAMEO concept. After considering input from experts, the automated translation seems to perform with elevated accuracy, yet its domain validity strongly depends on the HITL input from the previous stage. The high degree of precision in the automated translation task can also be explained by the simplicity of translating a single word relative to translating longer and more complex sentences (Brown, Pietra, Pietra, Mercer, and Mohanty 1992; Sudoh, Duh, Tsukada, Hiaro, and Nagata 2010). Also, notice that the automated translation from English into Spanish maintains the verb in its infinitive form. This facilitates the translation task as it does not require Spanish conjugation for different genders, persons, and tenses — a task that can be highly complex. In this way, the performance of WordNet in Spanish at this low level of complexity is not surprising.

In addition to validating the proposed translations, coders provided valuable additions to the verbs dictionary as indicated in Table 1. Coders added 2,719 spanish verbs across different CAMEO categories, which represent about 7.3% of the total list of actions.[6] The coder's feedback also contributed to expanding the number of CAMEO rules by 155.1%. Finally, human coders also expanded the original *rule synsets* list by 357%. The rule synsets are wild cards used to increase the range of combinations in CAMEO rules. For

[5]Notice that the system only enabled assessing the Spanish verb translation for English synsets previously marked as correct.

[6]Most added verbs were country-specific colloquial verbs describing particular actions. For example, the word "*encajuelar*" refers to leaving a dead body inside a car trunk. "*Encajuelar*" is specific to Mexico, while other Spanish-speaking countries use the term "*embaular*" to refer to the same action.

Table 1. Dictionary additions.

Item	Original	New	Percent increase
Verbs	37,440	2,719	7.3%
Rules	1,999	3,100	155.1%
Rule synsets	598	2,137	357.4%

example, the rule synset &CURRENCY contains a long list of currency names (such as Peso or Dollar) that can be inserted into any rule for efficiently linking a verb to any currency.

Given the range of actions considered under CAMEO's 20 categories, it is useful to identify the distribution of verdicts across different categories in order to gain insights about the extent of behaviors covered in the dictionary. Figure 5 shows the percentage of verdicts (correct, ambiguous, and incorrect) by category. Each category comprises multiple related words in English, as well as their definitions and proposed examples extracted from WordNet (Step 4 in the translation process). On average, coders issued 2,053 verdicts across categories. The CAMEO codes that received most verdicts are 16: Fight with 7,494 verdicts (18.2% of the total), code 4: Yield with 4,678 verdicts (11.4%), and 7: Make Public Statements with 3,840 verdicts (9.4%). In contrast, code 17: Engage in Unconventional Mass Violence received the least verdicts in only 183 occasions (0.4% of the total), followed by 15: Threaten with 452 verdicts (1.1%), and 14: Consult with 737 verdicts (1.8%).

The analysis of verdicts also reveals variations in the difficulty of matching English synsets to specific CAMEO code meanings. Figure 6 shows the

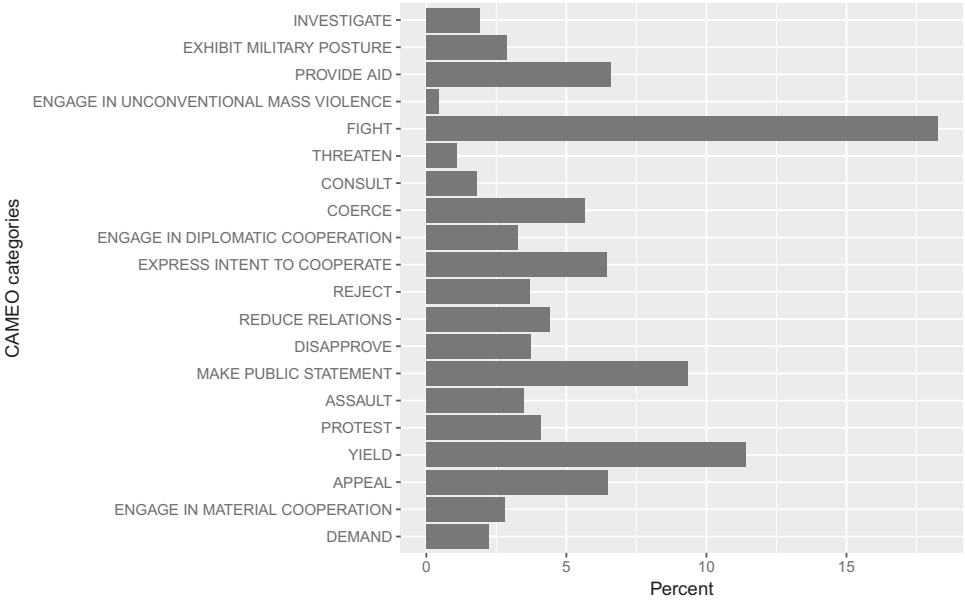


Figure 5. Percentage of english synset verdicts by CAMEO category.

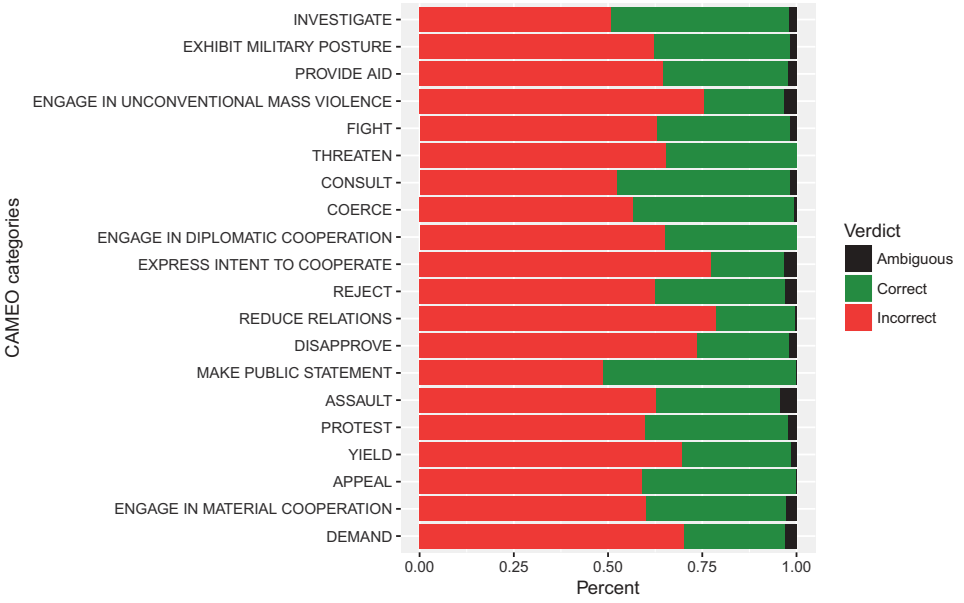


Figure 6. Average english synset verdict by CAMEO Category.

average verdict categories across different CAMEO codes. In general, the average proportion of correct verdicts is 34.3%, whereas 63.0% of the synsets were coded as incorrect, and the 1.8% remaining were ambiguous. According to Figure 6, the categories with the highest rates of match between the English synset and the CAMEO code are 7: Make Public Statement with 51.1% correct verdicts, followed by 14: Consult with 54.7%, and 13: Coerce with 42.6%. In contrast, the categories with the least percentage of correct matches between the English synset and the CAMEO meaning are 11: Express Intent to Cooperate with 19.4%, 9: Reduce Relations with 21.2%, and 17: Engage in Unconventional Mass Violence with 31.3%. In general, the homogeneity of the verdict distribution across CAMEO categories dissipates concerns of bias in favor or against any specific QuadClass sector of material/verbal-conflict/cooperation.

CAMEO VTA also provides valuable insights for intercoder reliability. The system tracks individual coder's verdicts and generates an overall sense of agreement among them. In general, there is a correlation factor of −0.96 between the average proportion of correct and incorrect verdicts across users. This indicates a high degree of agreement between coders in which they consistently identify English synsets as correct when they indeed correspond to the CAMEO concept, and consistently tag them as incorrect when their definition does not align with CAMEO's meaning. However, as Figure 7 indicates, assessing the match of synsets to CAMEO concepts is not straightforward as there are some instances of language ambivalence. The "Incorrect" and "Correct" dimensions of Figure 7 show a tight negative correspondence

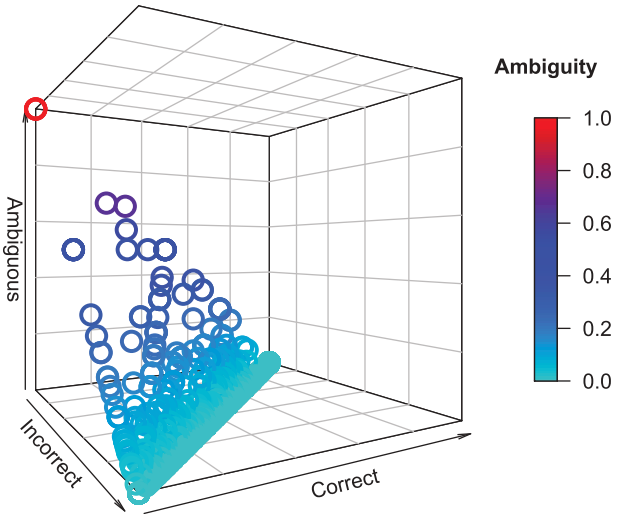


Figure 7. Intercoder verdict consistency.

between correct and incorrect synset verdicts. However, there are a few instances in the middle range in which coders are ambivalent in their collective judgments. These instances rise up along the "Ambiguous" dimension.[7] The synsets and their corresponding translations that were accepted into the Spanish CAMEO verb dictionary are those in which there is a rate of agreement of at least 75% among coders issuing a correct verdict.

Future Research

This research note presents CAMEO VTA as an effective application for translating CAMEO verb dictionaries to non-English languages in a systematic and encompassing way. This collaborative application relies on the strengths of automated translation systems while implementing a HITL approach to incorporate the expertise and nuance of human coders into the coding process. This application allows researchers to capitalize on the years of work spent building the CAMEO dictionaries which serve as the industry standard in computerized event coding. In addition, this technological development opens the possibility of expanding CAMEO's potential to process events in non-English languages in a systematic manner while maintaining comparability of code categories across languages. By enabling the capacity to process non-English languages, this research addresses the limitation of most computerized event coding projects that primarily rely on English text as their source of information. However, as this study shows, it is important to proceed carefully when translating CAMEO to other languages.

[7]The verb "to maintain" ("*mantener*") is an example of an ambiguous action. Without contextual information, it is difficult to identify its correspondence to a CAMEO category.

The application of CAMEO VTA for translating verbs into Spanish reveals the crucial role of human intervention in validating the suggestions made by automated protocols. As the performance metrics show, approximately two-thirds of the automated synset suggestions do not match the meaning of the behavioral category that the CAMEO dictionary aims to capture, and coders are largely in agreement about this assessment. This indicates that the input from expert coders is essential for ensuring the validity of the coding instruments.

The centrality of the human-in-the-loop in this application has direct implications for ongoing efforts of computerized event coding and for natural language processing more broadly. Despite the attractive possibilities that technological developments open for expanding the use of text-as-data in Political Science (see Cardie and Wilkerson 2008; de Vries et al. 2018; Grimmer and Stewart 2013; Mueller and Rauh 2016; Ulfedler 2013), domain-specific coding projects are still highly dependant on the input provided by human experts for two reasons. First, most event coding projects rely on unstructured text (for example, news stories) that confront the researchers with an enormous variation of grammatical and syntactical complexity. Given the multiple meanings words might have depending on the contextual information, the abstraction capacity of humans is crucial for identifying the correct meaning of words. Even if abstraction and meaning are comparable, there might be syntactical and grammatical differences across languages that make some words not directly commensurable. As Sartori (1970) noted, concepts may not travel coherently across languages and cultures. A clear indication of such nuances is the contribution of a considerable number of new verbs, rules, and rule synsets added by our coders. In many of these instances, coders took into consideration colloquialisms, journalism-specific jargon, and their knowledge of society and culture in different Spanish-speaking countries. Without this human input, current automated translation tools might suffer from limited validity. Of course, such concern might be lower if the research project relies on structured text (for example, government administrative records, or medical diagnoses). The reason is that the content of such structured text might have already a higher degree of validity when compared to structured text.

Second, event coding projects usually focus on analyzing a specific type of behavior, known as the research domain. For example, Gerner, Schrodt, and Yilmaz (2009) analyze the Israeli-Palestinian conflict, Osorio (2015) studies organized crime violence, and Hanna (2017) focuses on protest events. Different meanings of a word might gain higher relevance depending on their alignment to specific meanings of concepts of the domain of study. Human coders then serve as judges to designate which word meanings correspond to the concepts or applications relevant to the domain under study.

We expect future applications of CAMEO VTA to translate CAMEO verb categories to other non-English languages so researchers can gain more

accurate and fine-grained data about cooperative or conflictive behaviors of verbal or material nature in foreign locations, and offer evidence-based responses to pressing substantive questions. In addition to Spanish, the Open Event Data Alliance is also working on translating CAMEO to Arabic. We make CAMEO VTA available for other researchers to expand CAMEO's potential to other languages and move toward a multilanguage collaborative environment. Such multilingual approach would require testing the accuracy of different coding instrument using parallel corpora and universal dependencies.

Funding

This work was supported by the National Science Foundation [SBE-SMA-1539302].

ORCID

Javier Osorio http://orcid.org/0000-0002-4548-7794

References

Brandt, Patrick, John R. Freeman, and Philip A. Schrodt. (2011) Real Time, Time Series Forecasting of Inter- and Intra-State Political Conflict. *Conflict Management and Peace Science*, 28(1):41–64. doi:10.1177/0738894210388125

Brown, Peter, Stephen Della Pietra, Vincent Della Pietra, Robert Mercer, and Surya Mohanty. (1992) Dividing and Conquering Long Sentences in a Translation System. Paper presented at Speech and Natural Language: Proceedings of a Work-shop Held at Harriman, New York, February 23-26, 1992. pp. 267–271. Available at http://www.aclweb.org/anthology/H92-1053.

Cardie, Claire, and John Wilkerson. (2008) Text Annotation for Political Science Research. *Journal of Information Technology and Politics*, 5(1):1–6. doi:10.1080/19331680802149590

de Vries, Erick, Martijn Schoonvelde, and Gijs Schumaher. (2018) No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis*, 26(4):417–430. doi:10.1017/pan.2018.26

Fellbaum, Christiane. (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gerner, Deborah J., Philip A. Schrodt, and Ömür. Yilmaz. (2009) Conflict and Mediation Event Observations (CAMEO): An Event Data Framework for a Post Cold War World. In *International Conflict Mediation: New Approaches and Findings*, edited by Jacob Bercovitch and Gartner. Scott. Chap. 13, New York: Routledge, pp. 287–304.

Grimmer, Justin, and Brandon M. Stewart. (2013) Text as Data: The Promise and Pit-Falls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297. doi:10.1093/pan/mps028

Hanna, Alex. (2017) MPEDS: Automating the Generation of Protest Event Data. Available at https://osf.io/preprints/socarxiv/xuqmv.

Herkenrath, Mark, and Alex Knoll. (2011) Protest Events in International Press Coverage: An Empirical Critique of Cross-National Conflict Databases. *International Journal of Comparative Sociology*, 52(3):163–180. doi:10.1177/0020715211405417

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Brthard, and McClosky. David (2014) The Stanford CoreNLP Natural Language Processing Toolkit. Paper presented at the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, June 22-27.

Miller, George A. (1995) WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41. doi:10.1145/219717.219748

Mueller, Hannes, and Christopher Rauh. (2016) Reading between the Lines: Prediction of Political Violence Using Newspaper Text. Paper presented at the LSE Political Science and Political Economy Workshop, London, 2 February 2016.

O'Brien, Sean P. (2010) Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, 12 (1):87–104. doi:10.1111/j.1468-2486.2009.00914.x

Osorio, Javier. (2015) The Contagion of Drug Violence: Spatiotemporal Dynamics of the Mexican War on Drugs. *Journal of Conflict Resolution*, 59(8):1403–1432. doi:10.1177/0022002715587048

Osorio, Javier, and Alejandro Reyes. (2016) Supervised Event Coding from Text Written in Spanish: Introducing Eventus ID. *Social Science Computer Review*, 35(3):406–416. doi:10.1177/0894439315625475

Sartori, Giovanni. (1970) Concept Misformation in Comparative Politics. *The American Political Science Review*, 64(4):1033–1053. doi:10.2307/1958356

Schrodt, Philip A. (2012) Precedents, Progress, and Prospects in Political Event Data. *International Interactions*, 38(4):546–569. doi: 10.1080/03050629.2012.697430.

Schrodt, Philip A. (2001) Automated Coding of International Event Data Using Sparse Parsing Techniques. Paper presented at the annual meeting of the International Studies Association, Chicago, IL, February 20-24. Available at http://polmeth.wustl.edu/media/Paper/schro01b.pdf.

Schrodt, Philip A. (2006) Twenty Years of the Kansas Event Data System Project. *The Political Methodologist* 14(1):1–6.

Schrodt, Philip A. (2009) TABARI. Textual Analysis by Augmented Replacement Instructions. Available at http://eventdata.parusanalytics.com/software.dir/tabari.html.

Schrodt, Philip A., John Beieler, and Muhammed Idris. (2014) Three's a Charm?: Open Event Data Coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance. Paper presented at the annual meeting of the International Studies Association, Toronto, March 26-29. Available at http://parusanalytics.com/eventdata/papers.dir/Schrodt-Beieler-Idris-ISA14.pdf.

Schrodt, Philip A., and Van Brackel. David. (2013) Automated Coding of Political Event Data. In *Handbook of Computational Approaches to Counterterrorism*, edited by Devika Subramanian. New York: Springer, pp. 23–50.

Schrodt, Philip A., Shannon G. Davis, and Judith L. Weddle. (1994) Political Science: KEDS - A Program for the Machine Coding of Event Data. *Social Science Computer Review*, 12 (4):561–587. doi:10.1177/089443939401200408

Shellman, Stephen M., Clare Hatfield, and Maggie J. Mills. (2010) Disaggregating Actors in Intranational Conflict. *Journal of Peace Research*, 47(1):83–90. doi:10.1177/0022343309350029

Sudoh, Katsuhito, Kevin Duh, Hajime Tsukada, Tsutomu Hiaro, and Masaaki Nagata. (2010) Divide and Translate: Improving Long Distance Reordering in Statistical Machine Translation. Paper presented at the Joint Fifth Workshop on Statistical Machine

Translation and Metrics, Uppsala, Sweden, July, 2010. pp. 418–427. Available at https://dl.acm.org/citation.cfm?id=1868850.1868912.
Ulfedler, Jay. (2013) The Future of Political Science Just Showed Up. Available at https://dartthrowingchimp.wordpress.com/2013/04/10/the-future-of-political-science-just-showed-up/.

Appendix

Definitions:

- HITL: human-in-the-loop. Input from human experts.
- VTA: Verb Translation Application. A generic web-based platform to assist the translation of the CAMEO verbs dictionary into other languages.
- CAMEO: Conflict and Mediation Event Observations.
- Synset: A sets of cognitive synonyms.
- Related Words: a set of verbs that capture different aspects of the underlying concept of the referred CAMEO code.
- Ontology: is a set of classes organized into a network with arbitrarily many kinds of relationships. A key one is the subclass relationship which forms a taxonomy backbone of an ontology. The relationships themselves have properties that are used for inferencing. For example, if A is a sibling of B, we know that B is a sibling of A. But this symmetry does not hold for the brother relationship. The meaning of the links is formal and automated reasoning is supported.
- WordNet: the industry standard in online linguistics tools. It is a large lexical database of English grouping nouns, verbs, adjectives, and adverbs into cognitive synonyms sets (synsets), each expressing a distinct concept (Fellbaum 1998; Miller 1995).
- Rule Synset: wild cards used to expand the range of combinations in CAMEO rules.