RWTH-Phoenix: Analysis of the German Sign Language Corpus

Daniel Stein, Jens Forster, Uwe Zelle, Philippe Dreuw, and Hermann Ney

Human Language Technology and Pattern Recognition
RWTH Aachen University, Germany
surname@cs.rwth-aachen.de

Abstract

In this work, the recent additions to the RWTH-Phoenix corpus, a data collection of interpreted news announcement, are analysed. The corpus features videos, gloss annotation of German Sign Language and transcriptions of spoken German. The annotation procedure is reported, and the corpus statistics are discussed. We present automatic machine translation results for both directions, and discuss syntactically motivated enhancements.

1. Introduction

For data-driven automatical sign language processing, finding a suitable corpus is still one of the main obstacles. Most available data collections focus on linguistic issues and have a domain that is too broad to be suitable for these approaches. In (Bungeroth et al., 2006), the RWTH-Phoenix corpus was described, a collection of richly annotated video data from the domain of German weather forecasting. It includes a bilingual text-based sentence corpus and a collection of monolingual data of the German sentences. This domain was chosen since it is easily extendable, has a limited vocabulary and features real-life data rather than material made under lab conditions.

In this work, we are going to analyse the recent additions made to the existing corpus and its impact on the automatic machine translation. We are also applying some recent advancements in the field of statistical machine translations and analyse if they work on tiny data collections.

1.1. Related Work

Recently, a couple of other sign language data collections have been created. Based on their initial purpose, some of them have only limited usability to data-driven natural language processing techniques. Listed below are some of the larger efforts for European sign languages.

ECHO The European Cultural Heritage Online organization (ECHO)[1] published data collections for Swedish Sign Language, British Sign Language and Sign Language of the Netherlands. Their broad domain of children's fairy tales as well as poetry make them rather unsuitable for statistical methods. Another obstacle is the intensive usage of signed classifiers because of the rather visual topics.

Corpus NGT (Crasborn and Zwitserlood, 2008) present a data collection in the Sign Language of the Netherlands. It consists of recordings in the domain of fables, cartoon paraphrases, discussions on sign language and discussions on Deaf[2] issues. In the european funded Signspeak project[3], sentence-aligned translations into Spoken Dutch are currently ongoing.

ATIS In (Bungeroth et al., 2008), a corpus for English, German, Irish Sign Language, German Sign Language and South African Sign Language in the domain of the Air Travel Information System (ATIS) is given. With roughly 600 parallel sentences in total, it is small in size. However, being a multilingual data selection, it enables direct translation between sign languages.

Czech–Signed Speech In (Kanis and Müller, 2009), a data collection for Czech and Signed Czech is presented. Its domain is taken from transcribed train timetable dialogues and then translated by human experts. However, the actual translations are not in the Czech Sign Language spoken by the Deaf, but in an artifical language system strongly derived from spoken Czech. Explicit word alignments are made by human experts. Due to its nature, the authors are able achieve with very high performance scores.

1.2. Paper Structure

This paper is organised as follows. We analyse the current status of our data collection in Section 2., with special attention to the transcription process and the corpus statistics. In Section 3., the translation methods and results are presented, including syntactically motivated enhancements to the translation system. In Section 4., a summary and an outlook are given.

2. Corpus Analysis

The public broadcast channel "Phoenix" offers live interpretation into German Sign Language (DGS)[4] for the main evening broadcast news. Its videos are recorded automatically by our servers.

Since the last batch of recordings in 2005 (Bungeroth et al., 2006), the television program has changed in two important aspects. First, the format of the video is different: before, the news announcer was slightly distorted in perspective, and the signing interpreter was shown without a background of its own. Now, the broadcast channel shows

[1] http://www.let.kun.nl/sign-lang/echo

[2] Following common conventions, we denote the cultural group of deaf people with a capital "D"

[3] http://www.signspeak.eu/

[4] Deutsche Gebärdensprache

the original video in a smaller frame and places the signing interpreter in front of a grey background on the far right (cf. Figure 1). For machine translation it does not pose a problem since the algorithms only work on the transcriptions and not on the video signal.

As for the second major change in the data, the transcription of the audio material is no longer provided by the broadcast station. We therefore employ an automatic speech recognition system for the German audio data which transcribes the spoken words, and manually align the words to the annotated gloss sentences. For the weather forecast, the audio recognition word error rate is well below 5%, making the transcription quite convenient.

2.1. Quality and Usability

Although interpreted by bilingual experts, the translation into German Sign Language quality suffers from the recording situation: the interpreters have to listen to the text under real-time conditions and thus have to sign simultaneously without preparations. Due to the complex nature of official news announcements and the relative speed of the announcer, the signed sentences are still in German Sign Language but tend to have a slight bias towards the grammar structure of spoken German. Also, details are omitted in the signed sentences. For example, if the temperature for the region of Bavaria, the adjacent Austrian Alps and the river Donau is described in the weather forecast, the interpreter might refer more generally to the south of Germany without specifically naming the exact locations. Another typical omittance occurs when the announcement refers to specific wind velocities such as "schwach", "mäßig", and "frisch" (being a 3, 4 and 5 on the Beaufort scale, respectively), the interpreters typically only differentiate between a low and a high velocity.

The notion of a signed sentence is an active research topic in the linguistic community. Here, we take a rather pragmatic (and probably erroneous) approach and match the gloss output to the spoken German sentences, i.e. we split gloss sentences transcribed by our deaf colleague if their topic stretches over more than one German sentence. In a second-pass, we also omit all information in the spoken German sentences that are clearly not signed by the interpreter, but try to stay as close to the previous grammar structure as possible.

2.2. Notation

According to common conventions, glosses are generally written in upper case. Incorporations are treated as a single word, finger-spelled words and compound words are joined by a $+$. Dialectal forms are stored in a simple database so that they are mapped to the same word for translation but appear differently for the recognition (e.g. "WOMAN1" for the Bavarian sign for woman and "WOMAN2" for the dialectal form used in the northern part of Germany). If a sign is repeated fast and without a specific number, a double $+$ is written at the end of the sign (e.g. "ASK++", which translates to *enquire* rather than *asking*). If a sign is repeated a specific number of times to mark multiple occurences, they are denoted separately (e.g. two groups of clouds are denoted as "CLOUD CLOUD"). Additional information that carries crucial semantic information is denoted as:

loc: for a specific location with a spatial reference (e.g. "loc:coast" for the coast in the northern part of Germany, but also "loc:from_north_to_south" for a southward movement)

mb: mouthing that is important to discriminate the word meaning, (e.g. "RIVER-(mb:rhein)" and "RIVER-(mb:donau)" for the different rivers which have the same manual movement)

Apart from this, we annotate hand movement not related to a signed word.i <ON>, <OFF> is used for signing onset and offset, <PU> is a palm-up gesture, and <EMP> marks emphatic movement that is not a sign (e.g. when the interpreter is shrugging the shoulder). For the translation experiments below, we treated the mouthing and location information as normal words.

2.3. Annotation

For the annotation, we made use of the free ELAN tool developed at the Max Planck Institute for Psycholinguistics in Nijmegen[5]. Start and end times are marked on a sentence level rather than on the word level. Both left-hand and right-hand movements are kept track of independently.

Our annotator is congenitally deaf and has worked in research fields regarding sign language for over a decade, but had no previous annotation experiences. According to his feedback, it took him about two weeks to get accustomed to the annotation tool. For the first two month working on the recordings, there were various questions coming up about the annotation procedure, namely for such effects as dialects, synonyms, classifiers, left-hand/right-hand issues which were discussed in his mother tongue with interpreters. At first, it took him 4 hours for one weather forecast of roughly one minute. After two months, he was able to finish three videos in the same time amount. For the whole news announcement, which has a basically unlimited domain and runs for 15 minutes, it takes him about 24 working hours to transcribe it.

2.4. Corpus Progression

In an ongoing process, the corpus has recently been extended with additional material. For the transcription of the glosses and their translation into spoken German, they blend in with the old annotations and can be used together for statistical machine translation. So far, 43 new videos were added to the existing 78 videos.

Comparing the corpus statistics with other small-sized data selections, the domain seems to be suitable. For example, the Chinese-English task of the International Workshop on Spoken Language Technology (IWSLT)[6], is a selection of parallel sentences in the domain of travel and booking information, has 22 K training sentences, with a token-type ratio of 18.8 for Chinese and 27.5 for English. Compared to our corpus, we currently have a total of 2.7 K training sentences and already approach a type-token ratio of around

[5] http://www.lat-mpi.eu/tools/elan/

[6] http://mastarpj.nict.go.jp/IWSLT2009/

Figure 1: Old and new television format used in the Phoenix television channel

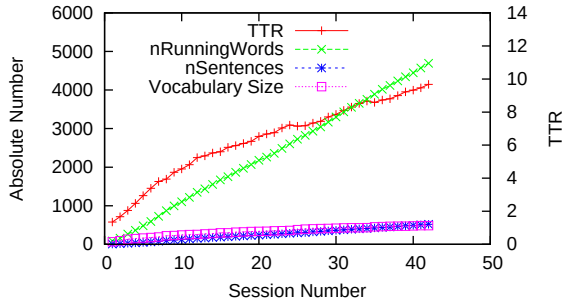

Figure 2: Number of sentences, vocabulary size, type-token ratio, for the newly annotated data

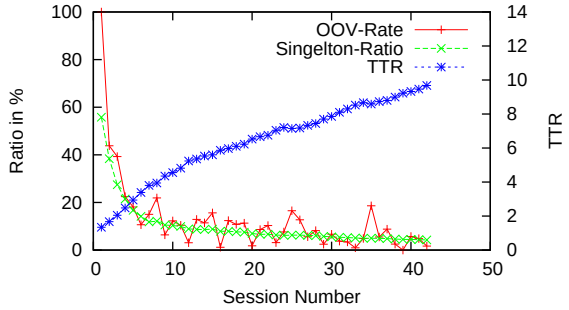

Figure 3: Out of vocabulary and singletons for the newly annotated data

10 (Figure 2 and 3) after 40 sessions. The singleton ratio is about 40% for both languages in IWSLT, while ours goes quickly below 20% and stays there. The peaks in singletons and out-of-vocabulary ratios can mostly be attributed to time-specific terms like the easter season or certain places where weather phenomena occur in a certain week. Since these words tend to occur often in consecutive sessions, the singleton ratio typically drops fast. For a complete corpus overview, see Table 1.

3. Translation

We use an in-house statistical translation system similar to (Chiang, 2005). It is able to process hierarchical phrases in a context-free grammar with a variation of the CYK algorithm. For a given sentence f, the best translation $\hat{e}$ is chosen as the target sentence e that maximizes the sum over m different models h_m, scaled by the factors λ_m:

$$\hat{e} := \underset{e}{\operatorname{argmax}} \left(\sum_m \lambda_m h_m(e, f) \right) . \quad (1)$$

The alignment is created for both translation directions with GIZA++[7] and merged with a variation of the grow-diag-final algorithm. We employ a trigram language model using modified Kneser-Ney discounting which is trained with the SRI toolkit[8]. The scaling factors of the log-linear model are optimized on the development set with Och's Minimum Error Rate Training (Och, 2003), which is a variation of Powell's method working on n-best translations. The resulting factors are then used to translate the test set. For automatic error measures, we use the Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2001), which is based on n-gram precision and has a brevity penalty for sentences that are too short. Further, we use the Translation Error Rate (TER) (Snover et al., 2006), which is similar to the Levenshtein distance but allows for shifts of word blocks. Note that BLEU is better if higher and TER is better if lower.

In order to enhance the statistic reliability of the results, we opted to increase the number of sentences withheld from the training material for development and test set to 20% rather than 10% in our previous publications. Further, cross-validation has been carried out, taking three different splits of the data into the training, development and testing set, with completely independent alignment creation, language model and optimization. The results between the splits are not comparable in this way, but a consistent improvement in all splits backs up the usefulness of the applied method.

[7] http://www.htlpr.rwth-aachen.de/~och/software/GIZA++.html
[8] http://www-speech.sri.com/projects/srilm/

		DGS	German	preprocessed German
train	#sentences	2711		
	#running words	15499	21679	22891
	vocabulary size	916	1476	1180
	#singletons	337	633	434
dev	#sentences	338		
	#running words	1924	2689	2832
	#OOVs	33	65	50
train	#sentences	338		
	#running words	1750	2629	2773
	#OOVs	48	49	32

Table 1: Corpus overview of the RWTH-Phoenix corpus for one specific split of the data. The numbers are similar for the other two splits. The preprocessing of German is explained in Section 3.1.

input	Wettervorhersage
	wetten–er–Vorhersage wettern–Vorhersage Wetter–Vorhersage ...
output	wettern Vorhersage

Table 2: Different breaking points proposed in Morphisto for the German word "Wettervorhersage" (english: weather forecasting). The last one is correct, but the second is taken in our heuristic.

3.1. German to Glosses

Sign languages lack a formal written form universally accepted by the Deaf. Thus, gloss annotations are typically only employed by linguistic experts, but they can be used to feed avatars with signing input. Being single-reference experiments, the quality of the output is reasonable but not without flaws. Looking at the examples in Table 3, we can see that the translation system was able to come up with some of the typical reorderings taking place in the grammar of the two languages, but failing to translate words that are highly flexed in German and thus lead to data sparseness problems.

We therefore reduced the morphologic complexity of the German source language by automatic means. To achieve this, we parsed the data with a morpho-syntactic analysis tool before the actual translation. The freely available tool Morphisto[9] is a finite-state transducer with a large database of German, accurately reporting part-of-speech tags, gender, casus and possible split points for large compound words. However, if ambiguous it does not provide probability scores for the various possible parsings. We therefore opted to always take the entry consisting of the fewest split points possible (cf. Table 2). By doing so, we reduce all words to their stem form and split large words automatically. In (Stein et al., 2006), it was already shown that these methods help enhance the translation quality.

In Figure 4, an example for an improvement in alignment quality is given. In Table 5, the results for this task are presented.

3.2. Glosses to German

This translation direction is more challenging since the German announcements often appear to be more variated and even lyrical in nature. Even though the interpreter always speaks of a clear sky during the night ("HEUTE NACHT KLAR"), the announcer will sometimes refer to the dissolving of the clouds, a clear sky or the sparkling of the stars. We are not able to preprocess the input automatically since no morpho-syntactic parser for the glosses exist, and a reduction of the target language complexity during translation would require a rather sophisticated post-processing that possibly introduces further errors.

[9] http://code.google.com/p/morphisto/

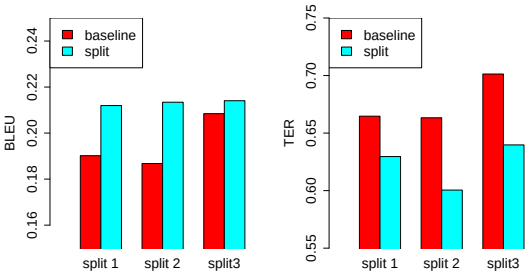

Figure 5: BLEU and TER results for German to German Sign Language Translation on three different test sets

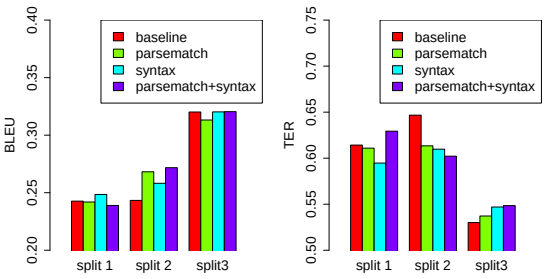

Figure 6: BLEU and TER results for German Sign Language to German Translation on three different test sets

However, we can make use of some syntactic analysis of the target language and enforce the structure of the German grammar onto our decoder. In this work, we opted for two methods. The first measures the compatibility of the phrases with a node in a deep syntactic tree, preferring complete sub-sentence structures such as noun phrases or verb phrases. If the target phrase does not match a node, we take the minimal amount of words needed to reach a fitting node as penalty, similar to (Vilar et al., 2008). We denote these experiments as *parsematch*.

Also, we employ soft syntactic features as in (Venugopal et al., 2009). With this, we replace the generic non-terminal label used in common hierarchical decoding and replace it

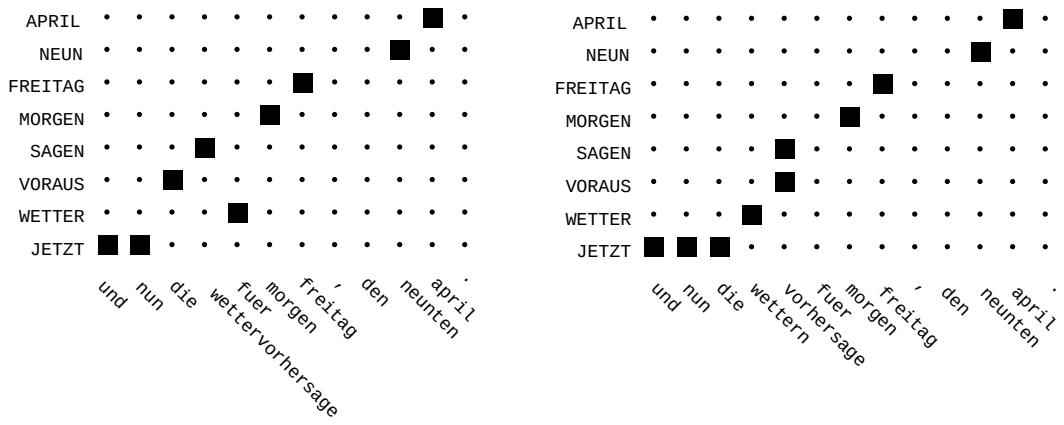

Figure 4: Alignment before and after splitting. The left one is more accurate.

source	Im Norden fällt etwas Regen bei stürmischem Wind .
baseline translation	NORDEN BISSCHEN REGEN HIER unknown_stürmischem
split translation	NORDEN BISSCHEN REGEN STURM-emp
reference	NORDEN WENIG REGEN STURM-emp
source	Diese Regenwolken ziehen heute Nacht aus Frankreich heran .
baseline translation	IX REGEN ZIEHEN HEUTE NACHT FRANKREICH
split translation	AUCH REGEN WOLKE ZIEHEN HEUTE NACHT FRANKREICH ZIEHEN-(loc:nach_mitte)
reference	REGEN WOLKE ZIEHEN++ FRANKREICH IX WOLKEN ZIEHEN-(loc:nach_mitte)

Table 3: Translation examples for German to German Sign Language

with phrase tags from the syntactic parser. Thus, we now have a variety of 65 non-terminals and see if a new translation matches the syntactic label that it tries to replace. This is denoted as *syntax*.

Note that we do not restrict the regular translation by doing so but merely offer another translation model to the log-linear model, thus theoretically allowing the decoding process to ignore it by setting the according scaling factor to 0. The parsing was done using the freely available Stanford parser[10].

In Figure 6, the results for this task are presented, with some examples in Table 4. While in general the BLEU score improved in all development optimizations, the results on our test test were not consistent. Possible reasons for this were the large number of labels that the Stanford parser produces, compared to the small data set. In a next step, we plan to reduce their number by means of automatic clustering. We also noted an increase in the TER score on some tasks, possibly by enforcing larger phrases with the syntactic models.

4. Conclusion

We presented and analysed the recent extensions to the signed weather forecasting corpus RWTH-Phoenix and tested various syntactically motivated methods to enhance the statistical machine translation on this task. It is currently one of the largest data collections for a natural sign language and designed for the needs of statistical translation and recognition. Great care has been taken to ensure that all the above mentioned methods and tools are freely available to the scientific community. Also, our complete hierarchical translation system will be released as open source in the near future.

The data collection is available upon request. We hope that the performance on this task can be taken for comparison and serve as a benchmark for other groups working in this field. As an outlook, we look forward to combine the recognition results made on the videos with our translation system. We also hope to further incorporate additional recognition input such as tracking of the hand position into our translation models.

[10]http://nlp.stanford.edu/software/lex-parser.shtml

5. Acknowledgments

This work has been partly funded by the European Community's Seventh Framework Programme (FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics - STREP) under grant agreement n° 231424.

6. References

Jan Bungeroth, Daniel Stein, Philippe Dreuw, Morteza Zahedi, and Hermann Ney. 2006. A german sign language corpus of the domain weather report. In *International Conference on Language Resources and Evaluation*, pages 2000–2003, Genoa, Italy, May.

Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. 2008. The atis sign language corpus. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.

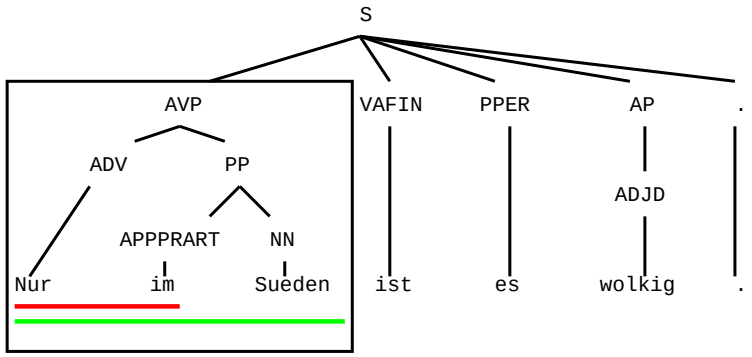

Figure 7: Stanford parsing of a German sentence. A phrase consisting of "Nur im" would be considered bad since it does not match a parse tree node. A good phrase would be "Nur im Sueden". For syntactic labeling, both phrases would be mapped to the same node, i.e. "AVP".

source	KALT LUFT MITTELMEER ZIEHEN-(loc:nach_mitte)
baseline	Kalte Mittelmeer Luft zu uns .
phrasematch	Kalte Luft Mittelmeer zieht zu uns .
reference	Kalte Mittelmeerluft zieht nach Deutschland .
source	DA TIEF DRUCK LOC REGION WESTEN DRUCK
baseline	Dort ein Tiefdruckgebiet liegt im Westen Hochdruckzone .
syntax	Weitgehend ein Tiefdruckgebiet im Nordwesten Tiefdruckgebiet bestimmt .
reference	Im Westen liegt ein Tiefdruckgebiet .

Table 4: Translation examples for German Sign Language to German

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270, Ann Arbor, Michigan, June.

Onno Crasborn and Inge Zwitserlood. 2008. The Corpus NGT: An Online Corpus for Professionals and Laymen. In Crasborn, Hanke, Efthimiou, Zwitserlood, and Thoutenhoofd, editors, *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages at LREC 2008*, pages 44–49, Paris. ELDA.

Jakub Kanis and Luděk Müller. 2009. Advances in czech - signed speech translation. In *Lecture Notes in Computer Science*, volume 5729, pages 48–55. Springer.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, September.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. pages 223–231, Cambridge, MA, August.

Daniel Stein, Jan Bungeroth, and Hermann Ney. 2006. Morpho-syntax based statistical methods for sign language translation. In *Conference of the European Association for Machine Translation*, pages 169–177, Oslo, Norway, June.

Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. 2009. Preference Grammars: Softening Syntactic Constraints to Improve Statistical Machine Translation. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–244, Boulder, Colorado, June.

David Vilar, Daniel Stein, and Hermann Ney. 2008. Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation. In *International Workshop on Spoken Language Translation*, pages 190–197, Waikiki, Hawaii, October.