# An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction

**Stefan Bott**
TALN Research Group
Universitat Pompeu Fabra
C/Tanger 122 - Barcelona - 08018
Spain
stefan.bott@upf.edu

**Horacio Saggion**
TALN Research Group
Universitat Pompeu Fabra
C/Tanger 122 - Barcelona - 08018
Spain
horacio.saggion@upf.edu

## Abstract

We present a method for the sentence-level alignment of short simplified text to the original text from which they were adapted. Our goal is to align a medium-sized corpus of parallel text, consisting of short news texts in Spanish with their simplified counterpart. No training data is available for this task, so we have to rely on unsupervised learning. In contrast to bilingual sentence alignment, in this task we can exploit the fact that the probability of sentence correspondence can be estimated from lexical similarity between sentences. We show that the algoithm employed performs better than a baseline which approaches the problem with a TF*IDF sentence similarity metric. The alignment algorithm is being used for the creation of a corpus for the study of text simplification in the Spanish language.

## 1 Introduction

Text simplification is the process of transforming a text into an equivalent which is more understandable for a target user. This simplification is beneficial for many groups of readers, such as language learners, elderly persons and people with other special reading and comprehension necessities. Simplified texts are characterized by a simple and direct style and a smaller vocabulary which substitutes infrequent and otherwise difficult words (such as long composite nouns, technical terms, neologisms and abstract concepts) by simpler corresponding expressions. Usually unnecessary details are omitted. Another characteristic trait of simplified texts is that usually only one main idea is expressed by a single sentence. This also means that in the simplification process complex sentences are often split into several smaller sentences.

The availability of a sentence-aligned corpus of original texts and their simplifications is of paramount importance for the study of simplification and for developing an automatic text simplification system. The different strategies that human editors employ to simplify texts are varied and have the effect that individual parts of the resulting text may either become shorter or longer than the original text. An editor may, for example, delete detailed information, making the text shorter. Or she may split complex sentences into various smaller sentences. As a result, simplified texts tend to become shorter than the source, but often the number of sentences increases. Not all of the information presented in the original needs to be preserved but in general all of the information in the simplified text stems from the source text.

The need to align parallel texts arises from a larger need to create a medium size corpus which will allow the study of the editing process of simplifying text, as well as to serve as a gold standard to evaluate a text simplification system.

Sentence alignment for simplified texts is related to, but different from, the alignment of bilingual text and also from the alignment of summaries to an original text. Since the alignment of simplified sentences is a case of monolingual alignment the lexical similarity between two corresponding sentences can be taken as an indicator of correspondence.

This paper is organized as follows: Section 2 briefly introduces text simplification which contex-



*Workshop on Monolingual Text-To-Text Generation*, pages 20–26,

*Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 20–26,
Portland, Oregon, 24 June 2011. 

tualises this piece of research and Section 3 discusses some related work. In Section 4 we briefly describe the texts we are working with and in Section 5 we present the alignment algorithm. Section 6 presents the details of the experiment and its results. Finally, section 7 gives a concluding discussion and an outlook on future work.

## 2 Text Simplification

The simplification of written documents by humans has the objective of making texts more accessible to people with a linguistic handicap, however manual simplification of written documents is very expensive. If one considers people who cannot read documents with heavy information load or documents from authorities or governmental sources the percent of need for simplification is estimated at around 25% of the population, it is therefore of great importance to develop methods and tools to tackle this problem. Automatic text simplification, the task of transforming a given text into an "equivalent" which is less complex in vocabulary and form, aims at reducing the efforts and costs associated with human simplification. In addition to transforming texts into their simplification for human consumption, text simplification has other advantages since simpler texts can be processed more efficiently by different natural language processing processors such as parsers and used in applications such as machine translation, information extraction, question answering, and text summarization.

Early attempts to text simplification were based on rule-based methods where rules were designed following linguistic intuitions (Chandrasekar et al., 1996). Steps in the process included linguistic text analysis (including parsing) and pattern matching and transformation steps. Other computational models of text simplification included processes of analysis, transformation, and phrase re-generation (Siddharthan, 2002) also using rule-based techniques. In the PSET project (Carroll et al., 1998) the proposal is for a news simplification system for aphasic readers and particular attention is paid to linguistic phenomena such as passive constructions and coreference which are difficult to deal with by people with disabilities. The PorSimples project (Aluísio et al., 2008) has looked into simplification of the Portuguese language. The methodology consisted in the creation of a corpus of simplification at two different levels and on the use of the corpus to train a decision procedure for simplification based on linguistic features. Simplification decisions about whether to simplify a text or sentence have been studied following rule-based paradigms (Chandrasekar et al., 1996) or trainable systems (Petersen and Ostendorf, 2007) where a corpus of texts and their simplifications becomes necessary. Some resources are available for the English language such as parallel corpora created or studied in various projects (Barzilay and Elhadad, 2003; Feng et al., 2009; Petersen and Ostendorf, 2007; Quirk et al., 2004); however there is no parallel Spanish corpus available for research into text simplification. The algorithms to be presented here will be used to create such resource.

## 3 Related Work

The problem of sentence alignment was first tackled in the context of statistical machine translation. Gale and Church (1993) proposed a dynamic programming algorithm for the sentence-level alignment of translations that exploited two facts: the length of translated sentences roughly corresponds to the length of the original sentences and the sequence of sentences in translated text largely corresponds to the original order of sentences. With this simple approach they reached a high degree of accuracy.

Within the field of monolingual sentence alignment a large part of the work has concentrated on the alignment between text summaries and the source texts they summarize. Jing (2002) present an algorithm which aligns strings of words to pieces of text in an original document using a Hidden Markov Model. This approach is very specific to summary texts, concretely such summaries which have been produced by a "cut and paste" process. A work which is more closely related to our task is presented in Barzilay and Elhadad (2003). They carried out an experiment on two different versions of the *Encyclopedia Britannica* (the regular version and the *Britannica Elementary*) and aligned sentences in a four-step procedure: They clustered paragraphs into 'topic' groups, then they trained a binary classifier (aligned or not aligned) for paragraph pairs

on a handcrafted set of sentence alignments. After that they grouped all paragraphs of unseen text pairs into the same topic clusters as in the first step and aligned the texts on the paragraph level, allowing for multiple matches. Finally they aligned the sentences within the already aligned paragraphs. Their similarity measure, both for paragraphs and sentences, was based on cosine distance of word overlap. Nelken and Shieber (2006) improve over Barzilay and Elhadad's work: They use the same data set, but they base their similarity measure for aligning sentences on a TF*IDF score. Although this score can be obtained without any training, they apply logistic regression on these scores and train two parameters of this regression model on the training data. Both of these approaches can be tuned by parameter settings, which results in a trade-off between precision and recall. Barzilay and Elhadad report a precision of 76.9% when the recall reaches 55.8%. Nelken and Shieber raise this value to 83.1% with the same recall level and show that TF*IDF is a much better sentence similarity measure. Zhu et al. (2010) even report a precision of 91.3% (at the same fixed recall value of 55.8%) for the alignment of simple English Wikipedia articles to the English Wikipedia counterparts using Nelken and Shieber's TF*IDF score, but their alignment was part of a larger problem setting and they do not discuss further details.

We consider that our task is not directly comparable to this previous work: the texts we are working with are direct simplifications of the source texts. So we can assume that all information in the simplified text must stem from the original text. In addition we can make the simplifying assumption that there are one-to-many, but no many-to-one relations between source sentences and simplified sentences, a simplification which largely holds for our corpus. This means that all target sentences must find at least one alignment to a source sentence, but not vice versa. Nelken and Shieber make the interesting observation that dynamic programming, as used by Gale and Church (1991) fails to work in the monolingual case. Their test data consisted of pairs of encyclopedia articles which presented a large intersection of factual information, but which was not necessarily presented in the same order. The corpus we are working with, however, largely preserves the order in which information is presented.

## 4 Dataset

We are working with a corpus of 200 news articles in Spanish covering the following topics: National News, Society, International News and Culture. Each of the texts is being adapted by the DILES Research Group from Universidad Autónoma de Madrid (Anula, 2007). Original and adapted examples of texts in Spanish can be seen in Figure 1 (the texts are adaptations carried out by DILES for Revista "La Plaza"). The texts are being processed using part-of-speech tagging, named entity recognition, and parsing in order to create an automatically annotated corpus. The bi-texts are first aligned using the tools to be described in this paper and then post-edited with the help of a bi-text editor provided in the GATE framework (Cunningham et al., 2002). Figure 2 shows the texts in the alignment editor. This tool is however insufficient for our purposes since it does not provide mechanisms for uploading the alignments produced outside the GATE framework and for producing stand-alone versions of the bi-texts; we have therefore extended the functionalities of the tool for the purpose of corpus creation.

## 5 Algorithm

Our algorithm is based on two intuitions about simplified texts (as found in our corpus): As repeatedly observed sentences in simplified texts use similar words to those in the original sentences that they stem from (even if some of the words may have undergone lexical simplification). The second observation is very specific to our data: the order in which information is presented in simplified texts roughly corresponds to the order of the information in the source text. So sentences which are close to each other in simplified texts correspond to original sentences which are also close to each other in the source text. In many cases, two adjacent simplified sentences even correspond to one single sentence in the source text. This leads us to apply a simple Hidden Markov Model which allows for a sequential classification.

Firstly, we define an alignment as a pair of sentences as

$$\langle source\_sent_i, target\_sent_j \rangle,$$

| Original Text | Adapted Text |
|---|---|
| Un Plan Global desde tu hogar<br>El Programa GAP (Global Action Plan) es una iniciativa que se desarrolla en distintos países y que pretende disminuir las emisiones de CO2, principales causantes del cambio climático y avanzar hacia hábitos más sostenibles en aspectos como el consumo de agua y energía, la movilidad o la gestión de los residuos domésticos.<br>San Sebastián de los Reyes se ha adherido a este Programa.<br>Toda la información disponible para actuar desde el hogar en la construcción de un mundo más sostenible se puede encontrar en ssreyes.org o programagap.es. | Un Plan Global desde tu hogar<br>San Sebastián de los Reyes se ha unido al Plan de Acción Global (GAP).<br>El Plan es una iniciativa para luchar contra el cambio climático desde tu casa.<br>Los objetivos del Plan son:<br>Disminuir nuestros gastos domésticos de agua y energía.<br>Reducir los efectos dañinos que producimos en el planeta con nuestros residuos.<br>Mejorar la calidad de vida de nuestra ciudad.<br>Tienes más información en ssreyes.org y en programagap.es.<br>Apúntate al programa GAP y descárgate los manuales con las propuestas para conservar el planeta. |

Figure 1: Original Full Document and Easy-to-Read Version

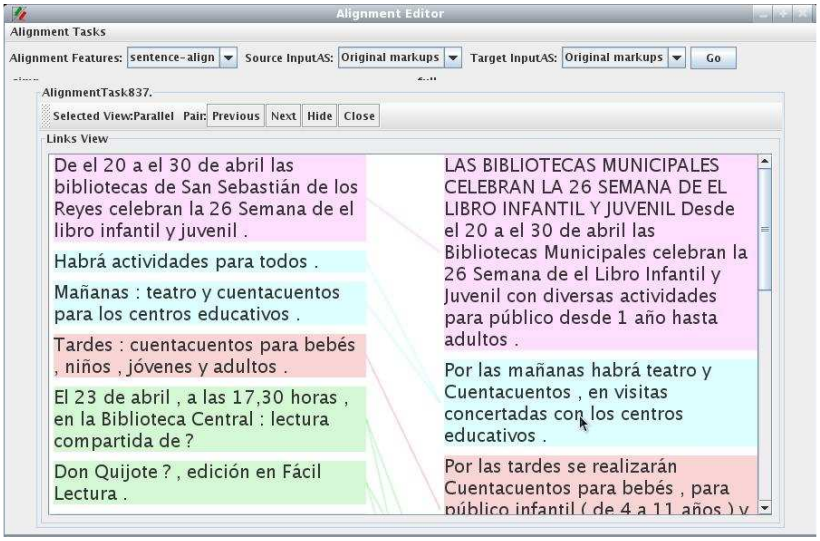


Figure 2: The Alignment Editor with Text and Adaptation

where a target sentence belongs to the simplified text and the source sentence belongs to the original sentence. Applying standard Bayesian decomposition, the probability of an alignment to a given target text can be calculated as follows:

$$P(align_1^n|target\_sent_1^m) = \frac{P(align_1^n)P(target\_sent_1^m|align_1^n)}{P(target\_sent_1^m)}$$

Since $P(target\_sent_1^m)$ is constant we can calculate the most probable alignment sequence $\widehat{align}$ as follows:

$$\widehat{align} = \textit{arg max} P(align_1^n)\ P(target\_sent_1^m|align_1^n) = \textit{arg max} \prod_{i=1}^{n} P(align_{i,j})P(target\_sent_j|align_{i,j})$$

This leaves us with two measures: a measure of sentence similarity (the probability of alignment proper) and a measure of consistency, under the assumption that a consistent simplified text presents the information in the same order as it is presented in the source text. In order to determine $\widehat{align}$, we apply the *Viterbi* algorithm (Viterbi, 1967).

Sentence similarity can be calculated as follows:

$$P(word_1^l|target\_sent_j) = \prod_{k=1}^{l} \frac{P(target\_sent_j|word_k)P(target\_sent_j)}{P(word_k)}$$

where $word_1^l$ is the sequence of words in the source sentence $i$ and $l$ is the length of sentence $i$.

This similarity measure is different from both word overlap cosine distance and TF*IDF. It is, however, similar to TF*IDF in that it penalizes

words which are frequent in the source text and boosts the score for less frequent words. In addition we eliminated a short list of stopwords from the calculation, but this has no significant effect on the general performance.

Note that $P(word_k)$ may correspond to a MLE of 0 since simplified texts often use different (and simpler) words and add connectors, conjunctions and the like. For this reason we have to recalculate $P(word_k)$ according to a distortion probability $\alpha$. Distortion is taken here as the process of word insertion or lexical changes. $\alpha$ is a small constant, which could be determined empirically, but since no training data is available we estimated $\alpha$ for our experiment and set it by hand to a value of 0.0075. Even if we had to estimate $\alpha$ we found that the performance of the system is robust regarding its value: even for unrealistic values like 0.0001 and 0.1 the performance only drops by two percent points.

$$P(word_k|distortion) = \\ (1-\alpha)P(word_k) + \alpha(1 - P(word_k))$$

For the consistency measure we made the Markov assumption that each alignment $align_{i,j}$ only depends on the proceeding alignment $align_{i-1,j'}$. We assume that this is the probability of a distance *d* between the corresponding sentences of $source\_sent_{i-1}$ and $source\_sent_i$, i.e. $P(source\_sent_i|align_{i-1,j-k})$ for each possible jump distance $k$. Since long jumps are relatively rare, we used a normalized even probability distribution for all jump lengths greater than 2 and smaller than -1.

Since we have no training data, we have to initially set these probabilities by hand. We do this by assuming that all jump distances *k* in the range between -1 and 2 are distributed evenly and larger jump distances have an accumulated probability mass corresponding to one of the local jump distances. Although this model for sentence transitions is apparently oversimplistic and gives a very bad estimate for each $P(source\_sent_i|align_{i-1,j-k})$, the probabilities for $P(align_1^n)$ give a counterweight to these bad estimates. What we can expect is, that after running the aligner once, using very unreliable transitions probability estimates, the output of the aligner is a set of alignments with an implicit alignment sequence. Taking this alignment sequence, we can calculate new maximum likelihood estimates for each jump distance $P(source\_sent_i|align_{i-1,j-k})$ again, and we can expect that these new estimates are much better than the original ones.

For this reason we apply the Viterbi classifier iteratively: The first iteration employs the hand set values. Then we run the classifier and determine the values for $P(source\_sent_i|align_{i-1,j-k})$ on its output. Then we run the classifier again, with the new model and so on. Interestingly values for $P(source\_sent_i|align_{i-1,j-k})$ emerge after as little as two iterations. After the first iteration, precision already lies only 1.2 percent points and recall 1.3 points below the stable values. We will comment on this finding in Section 7.

## 6 Experiment and Results

Our goal is to align a larger corpus of Spanish short news texts with their simplified counterparts. At the moment, however, we only have a small sample of this corpus available. The size of this corpus sample is 1840 words of simplified text (145 sentences) which correspond to 2456 (110 sentences) of source text. We manually created a gold standard which includes all the correct alignments between simplified and source sentences. The results of the classifier were calculated against this gold standard.

As a baseline we used a TF*IDF score based method which chooses for each sentence in the simplified text the sentence with the minimal word vector distance. The procedure is as follows: each sentence in the original and simplified document is represented in the vector space model using a term vector (Saggion, 2008). Each term (e.g. token) is weithed using as TF the frequency of the term in the document and IDF = $log(N + 1/M_t + 1)$ where $M_t$ is the number of sentences [1] containing $t$ and $N$ is the number of sentences in the corpus (counts are obtained from the set of documents to align). As similarity metric between vectors we use the cosine of the angle between the two vectors given in the following formula:

[1] The relevant unit for the calculation of IDF (the D in IDF) here is the sentence, not the document as in information retrieval.

$$\text{cosine}(s_1, s_2) = \frac{\sum_{i=1}^{n} w_{i,s_1} * w_{i,s_2}}{\sqrt{\sum_{i=1}^{n} (w_{i,s_1})^2} * \sqrt{\sum_{i=1}^{n} (w_{i,s_2})^2}}$$

Here $s_1$ and $s_2$ are the sentence vectors and $w_{i,s_k}$ is the weight of term i in sentence $s_k$. We align all simplified sentences (i.e. for the time being no cutoff has been used to identify new material in the simplified text).

For the calculation of the first baseline we calculate IDF over the sentences in whole corpus. Nelken and Shieber (2006) argue that that the relevant unit for this calculation should be each document for the following reason: Some words are much more frequent in some texts than they are in others. For example the word *unicorn* is relatively infrequent in English and it it may also be infrequent in a given colletion of texts. So this word is highly discriminative and it's IDF will be relatively high. In a specific text about imagenary creatures, however, the same word *unicorn* may be much more frequent and hence it's discrimiative power is much lower. For this reason we calcuated a second baseline, where we calculate the IDF only on the sentences of the relevanct texts.

Results of aligning all sentences in our sample corpus using both the baseline and the HMM algorithms are given in Table 6.

| | precision | recall |
|---|---|---|
| HMM aligner | 82.4% | 80.9% |
| alignment only | 81.13% | 79.63% |
| TF*IDF + transitions | 76.1% | 73.5% |
| TF*IDF (document) | 75.47% | 74.07% |
| TF*IDF (full corpus) | 62.2% | 61.1% |

If we compare these results to those presented by Nelken and Shieber (2006), we can observe that we obtain a comparable precision, but the recall improves dramatically from 55.8% (with their specific feature setting) to 82.4%. Our TF*IDF baselines are not directly comparable comparable to Nelken and Shieber's results. The reason why we cannot compare our results directly is that Nelken and Shieber use supervised learning in order to optimize the transformation of TF*IDF scores into probabilities and we had no training data available.

We included the additional scores for our system, when no transition probabilities are included in the calculation of the optimal alignment sequence and the score comes only from the probabilies of our clalculation of lexical similarity between sentences (*alignment only*). These scores show that a large part of the good performance comes from lexical similarity and sequencial classification only give an additional final boost, a fact which was already observed by Nelken and Shieber. We also attribute the fact that the system alrives at stable values after two iterations to the same efect: lexical similarity seems to have a much bigger effect on the general performance. Still our probability-based similarity meassure clearly outperforms the TF*IDF baselines.

## 7 Discussion and Outlook

We have argued above that our task is not directly comparable to Nelken and Shieber's alignment of two versions of Encyclopedia articles. First of all, the texts we are working with are simplified texts in a much stricter sense: they are the result of an editing process which turns a source text into a simplified version. This allows us to use sequential classification which is usually not successful for monolingual sentence alignment. This helps especially in the case of simplified sentences which have been largely re-written with simpler vocabulary. These cases would normally be hard to align correctly. Although it could be argued that the characteristics of such genuinely simplified text makes the alignment task somewhat easier, we would like to stress that the alignment method we present makes no use of any kind of training data, in contrast to Barzilay and Elhadad (2003) and, to a minor extent, Nelken and Shieber (2006).

Although we started out from a very specific need to align a corpus with reliably simplified news articles, we are confident that our approach can be applied in other circumstances. For future work we are planning to apply this algorithm in combination of a version of Barzilay and Elhadad's macro-alignment and use sequential classification only for the alignment of sentences within already aligned paragraphs. This would make our work directly comparable. We are also planning to test our algorithm, especially the sentence similarity measure it uses, on data which is similar the data Barzilay and Elhadad (and also Nelken and Shieber) used in their

experiment.

Finally, the alignment tool will be used to sentence-align a medium-sized parallel Spanish corpus of news and their adaptations that will be a much needed resource for the study of text simplification and other natural language processing applications. Since the size of the corpus we have available at the moment is relatively modest, we are also investigating alternative resources which could allow us to create a larger parallel corpus.

## Acknowledgments

We thank three anonymous reviewers for their comments and suggestions which help improve the final version of this paper. The research described in this paper arises from a Spanish research project called Simplext: An automatic system for text simplification (`http://www.simplext.es`). Simplext is led by Technosite and partially funded by the Ministry of Industry, Tourism and Trade of the Government of Spain, by means of the National Plan of Scientific Research, Development and Technological Innovation (I+D+i), within strategic Action of Telecommunications and Information Society (Avanza Competitiveness, with file number TSI-020302-2010-84). We thanks the Department of Information and Communication Technologies at UPF for their support. We are grateful to Programa Ramón y Cajal from Ministerio de Ciencia e Innovación, Spain.

## References

Sandra M. Aluísio, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick Galani Maziero, and Renata Pontin de Mattos Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *ACM Symposium on Document Engineering*, pages 240–248.

A. Anula. 2007. Tipos de textos, complejidad lingüística y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61.

Regina Barzilay and Noemi Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *In Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 25–32.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *In Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *COLING*, pages 1041–1044.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Lijun Feng, Noemie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *EACL*, pages 229–237.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*.

Hongyan Jing. 2002. Using hidden markov modeling to decompose human-written summaries. *Comput. Linguist.*, 28:527–543, December.

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *In 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *In Proc. of Workshop on Speech and Language Technology for Education*.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149.

H. Saggion. 2008. SUMMA: A Robust and Adaptable Summarization Tool. *Traitement Automatique des Langues*, 49(2):103–125.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *In LEC 02: Proceedings of the Language Engineering Conference (LEC02*, pages 64–71.

A. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics*, pages 1353–1361, Beijing, China, Aug.