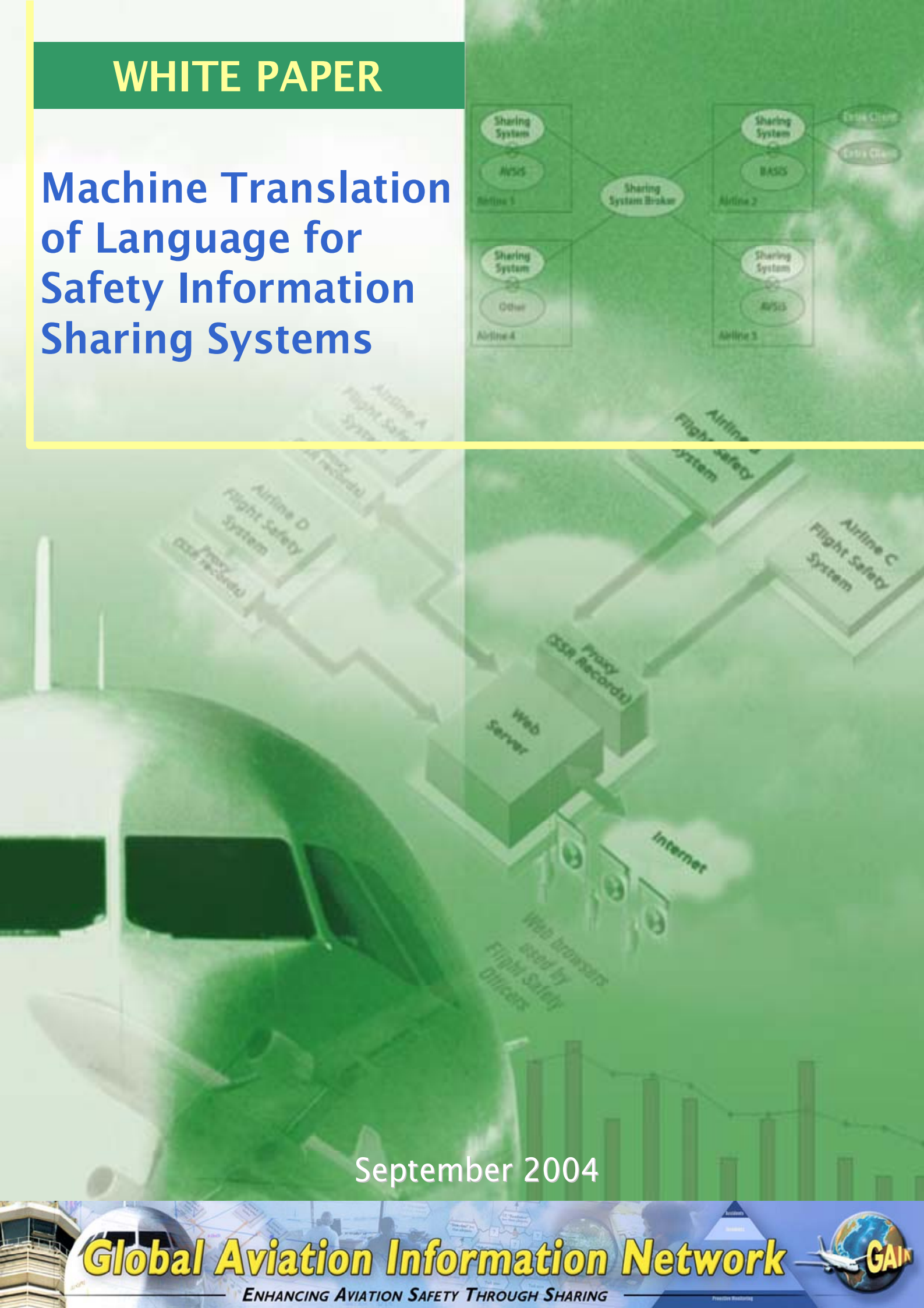
WHITE PAPER

# Machine Translation of Language for Safety Information Sharing Systems




September 2004

Global Aviation Information Network
ENHANCING AVIATION SAFETY THROUGH SHARING

## Disclaimers; Non-Endorsement

All data and information in this document are provided "as is," without any expressed or implied warranty of any kind, including as to the accuracy, completeness, currentness, noninfringement, merchantability, or fitness for any purpose.

The views and opinions expressed in this document do not necessarily reflect those of the Global Aviation Information Network or any of its participants, except as expressly indicated.

Reference in this document to any commercial product, process, or service by trade name, trademark, servicemark, manufacturer, or otherwise, does not constitute or imply any endorsement or recommendation by the Global Aviation Information Network or any of its participants of the product, process, or service.

## Notice of Right to Copy

# Background

The GAIN Action Plan outlines objective C.4 to promote international standardization of aviation safety data and information. Task C.4.c specifies Work Group C to document the capabilities of language translation tools that could be used by flight safety officers to share data in many languages. This report covers the research and analysis completed in response to this task. The document is not intended to represent a treatise on the subject of language machine translation. It is intended to deliver the findings and conclusions of the work group. The intent of this effort was to complete a survey of the available products and assess feasibility for implementation. This document presents a summary of the survey findings, cost analysis, and conclusions of the WG C.

# Translation

There are two fundamental methods of performing machine translation, direct and indirect.

**Direct** is the most simple of all techniques and requires limited computational resources. A simple comparison would be the use of a tourist's translation book. Each word is referenced in a dictionary and translated to the corresponding word in the destination language. Direct translation includes limited analysis of the sentence's syntactic structure.

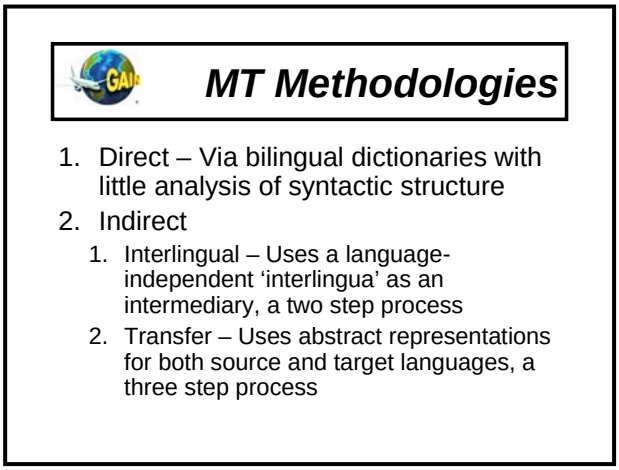



**Figure 1 Machine Translation Methodologies**

**Indirect** translation utilizes an intermediate resource in the translation process. **Interlingual** translation uses an independent Interlingua, which is an intermediate language using a standardized syntax. The original text is translated to the Interlingua that is subsequently translated to the target language. **Transfer** translation incorporates the use of abstract representations for the source and target languages. In both cases, analysis of the original sentence syntax is incorporated into the application of the intermediate resource.

There are no issues with respect to supporting aviation data for either direct or indirect methodologies. The benefits are in the accuracy of the indirect methods. Additional methods have been developed to improve translation quality beyond these basic categories.

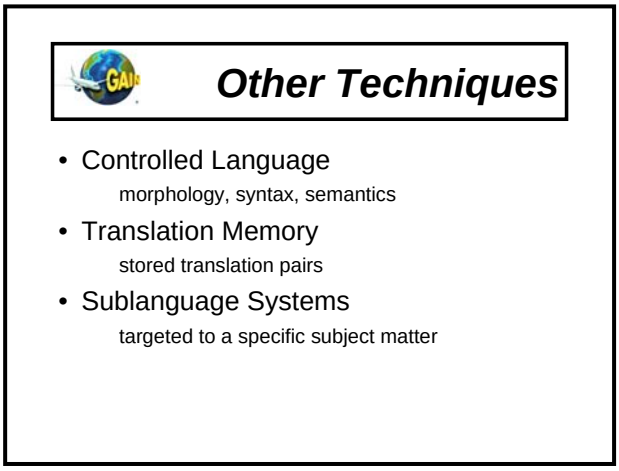



**Figure 2 Translation Techniques**

**Controlled Language** forces users to incorporate modifications to the syntax of the originating language. In this manner, the original text may not read as a normal passage by a native speaker of the subject language. The intent is to force a mandatory syntax that the translator expects and can then properly translate into the target language using proper syntax. This is not feasible in the aviation environment.

Sharing systems will employ a legacy data set that would not be revised to support the control language. Second, information sources will be global and in multiple languages. There cost and cultural challenges to train a new controlled grammar are infeasible.

**Translation Memory** retains knowledge of prior translations to reuse with each recurrence of the sentence or phrase. This methodology proves very efficient for repetitive translations of common passages (e.g., boilerplate). Additional value is found when the translation memory can be validated through human translation. It is undetermined if a translation memory would benefit translating safety narratives. Human validation would increase the value of store translation pairs, but it would most likely not be available due to cost.

**Sublanguages** are used when the subject matter involves specialized terminology and syntax. Sublanguage techniques would prove extremely valuable while translating words with aviation industry specific definitions.

The use of any particular translation methodology, including human translation, is based upon the use of the translation. This is called **Translation Demand**. There are four basis types of demand: dissemination, assimilation, interchange, and application.

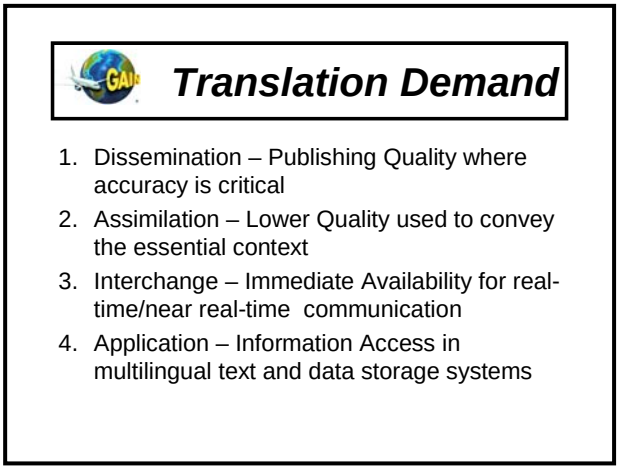



**Figure 3 Translation Service Demands**

**Dissemination** will distribute the translated passage to a wide audience and necessitates high accuracy. This is a publishing quality translation. In nearly all cases, the final review is accomplished by a human. The cost associated with these validated translations is prohibitive for shared safety narratives. Each participant could contribute several thousand narratives.

Sometimes, it is important to only assure that the general intent of a passage, within its intended context, is conveyed in the translation. These translations may include errors in grammar and/or phrasing. **Assimilation** allows a lower level of accuracy so long as it still provides the essential context. Human review may occur if necessary. Most flight safety officers that have been interviewed by GAIN WG B and/or WG C see the value in providing such context.

Modern internet applications allow individuals to communicate in two different languages via an "Internet Chat" style interface. The backend of the application provides translation services to translate outgoing messages to the target language and incoming messages to the local language. **Interchange** requires the translation of these messages at real-time or near real-time speed. The primary need is for speedy response. Accuracy is degraded due to limited time for syntax analysis. Human review is not possible given the need for real-time response. Translation used within a safety information sharing system is perceived to be from automated process to automated process.

Most computer applications in a multilingual environment do not directly interface with individual users. In these cases, automated processes identify new text entries needing translation. **Application** demand can take two forms. In the first type, new text values in a database are flagged; translation is accomplished into one or more target languages, and then stored for future access. This method supports the requirements of text mining in multiple languages. The second type performs on-demand translation. Text is not translated until

invoked by a user request. Since translations are not available for all text passages, text mining is not fully supported. Human participation in this environment is not possible. The cost of translating all content is prohibitive. In most cases, accuracy of the translations should be comparable to the requirements for Assimilation. Other constraints for the required level of accuracy are based upon systems functions (e.g., text mining function accuracy may be sensitive to translation accuracy).

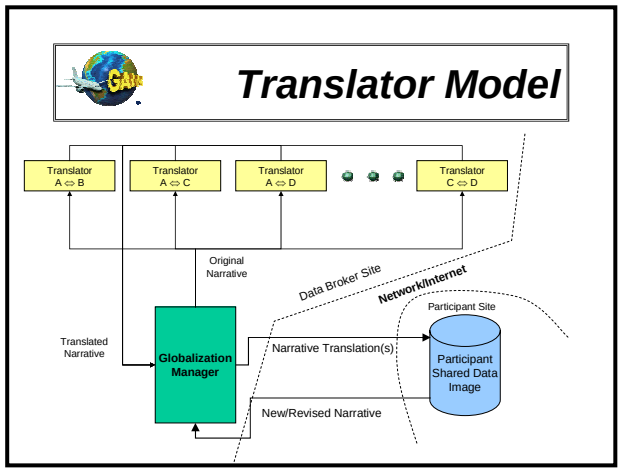



**Figure 4 Example Application Demand Topology**

The perceived demand from global sharing systems is for Applications. Participants in a trusted information sharing consortium will retain possession of their data. Translations would be accomplished as new or revised text is added to the data store. An application called the "Globalization Manager" tags new narrative, extracts the content from the data store, and submits the narrative to the bank of translation servers. The Globalization Manager and the Translator Servers can be either centralized resources for the entire system or localized for each participant. Centralized systems limit required asset procurement but increase network traffic. The Translator Servers would need to accommodate every possible language pair. Localized increases asset costs for each participant but reduces network traffic. The local Translator Servers would need to support only language pairs originating with the local language.

A summary of rolls for translation can be found in Figure 5

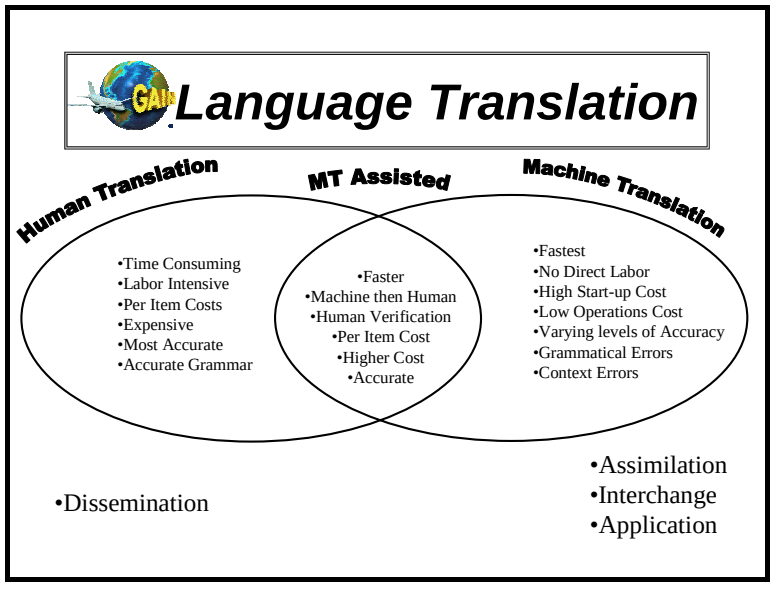



**Figure 5 Roles for Machine Translation**

## Cost and Availability

This analysis looked at the cost of machine translation packages for both stand alone systems (e.g., desktop PC applications) and server application and middleware installations.

Stand Alone applications would serve a purpose in supporting individual users to translate documents on an as needed basis. They will not support the implementation of multilingual data sharing systems.

This analysis looked at the availability of translation packages for language used by active GAIN participants. Prices, compiled for this analysis in March 2004, ranged from €163 to €797. For any given language pair, prices varied based upon the complexity of translation and the availability of comparable products. Translating across alphabets tended to be more expensive. For example, language pairs including Arabic or Japanese tended to be more expensive. For some language pairs, no known products exist. In some cases this may not be a problem. There most likely is a limited base of existing documents needing translation from Finnish to Japanese. But there is a gap in coverage for the six standard ICAO languages (English, French, Spanish, Russian, Arabic, and Chinese). Products for non-Latin alphabet languages are less common and more expensive.

**Stand Alone MT S/W**

| | English | French | Spanish | Arabic | Finnish | Japanese |
|---|---|---|---|---|---|---|
| English | | €403 MLTS | €163 LogoMedia | €480 MLTS | €329 TranSmart | €797 HonYaku |
| French | €403 MLTS | | €258 SysTran | €480 MLTS | N/A | €610 LogoVista |
| Spanish | €163 LogoMedia | €258 SysTran | | N/A | N/A | €610 LogoVista |
| Arabic | €480 MLTS | €480 MLTS | N/A | | N/A | N/A |
| Finnish | €329 TranSmart | N/A | N/A | N/A | | N/A |
| Japanese | €797 HonYaku | €610 LogoVista | €610 LogoVista | N/A | N/A | |

**Figure 6 Stand Alone Package Costs**

Server applications are significantly more expensive. Where some stand alone applications may include multiple language pairs, server applications are licensed for specific language pairs. Licenses are contracted for a defined term typically lasting from one to several years. Terms can vary but can be negotiated to include software updates, translation engine updates, vocabulary updates, technical support, and multiple CPUs.

Some of the more complex products will include specialized dictionaries supporting sublanguage systems, and specialized translation techniques. Systems utilizing statistical and pattern matching techniques are trained using pre-existing matched document pairs. The training can vary from constructing a translation memory to building probability models. Again, the demand type drives the required level of accuracy. Accuracy requirements establish whether to employ complex methods.

**Server S/W**

- By Language Pair
- Specialized Dictionaries
- Technical Support
- Software Updates
- License Terms
- €8000 +

**Figure 7 Server Translation Applications**

Services (e.g., software updates, technical support), proven accuracy ratings, and available competing products drive prices. This evaluation found that a one-year license may start around €8,000. High-end products with a multi-year or perpetual license can cost upwards of €80,000.

# Conclusions

GAIN Work Group C has concluded that it will not further pursue the development of translation tools for safety narratives. This decision was based upon concerns for cost, accuracy, and need.

A primary condition to sharing data amongst airline safety offices is that all data remain under the stewardship of the owners. This requires a distributed database architecture for any networked sharing system. The two methods for implementing machine translation services in a distributed architecture are as described above in the definition of **Application Demand**. Both methods prove expensive to implement.

A centralized translation process will require translation services for every possible language pair. Two languages will result in two language pairs supporting translation from language A to language B and the reverse. The number of language pairs that a centralized process requires is equal to $n(n-1)$, where $n$ is the number of languages used by the participants. Supporting the six ICAO standard languages will require 30 language pairs. In a distributed processing topology, each site would need to support up to 5 language pairs. The total number of language pairs would be the same as for a centralized process, but there could be multiple instances of each language pair. The benefit would be based upon the number of participants and the number of languages. The primary difference would be in the number of translation and globalization manager servers needed to support the system. In either case, the resulting cost is prohibitive.

Machine translation is a developing technology. Products claim proven accuracy rates to warrant their price. This evaluation did not conduct a thorough survey to determine a correlation between price and accuracy. Simple evaluations and product literature research did show that the more accurate translators tend to be more expensive. The WG C believes that given a realistic budget for machine translation products could not support the cost of products with the minimum acceptable accuracy.

ICAO via Annexes 1 and 11 mandates the use of English in radio communications. This primarily affects ATM services and airmen (i.e., pilots). ICAO Annex 13 establishes requirements for accident and incident investigation. Although it does not specifically state the required language for final reports, the annex does state that one of the six official ICAO languages be used for implementation. This does not place a direct requirement on airlines to record safety event information in one of the ICAO languages. However, most airline personnel are becoming fluent in English as the airlines prepare to comply with the ICAO mandate for language proficiency. Therefore, WG C concluded the cost and effort to implement a viable machine translation solution is unnecessary.

***PRINTING COURTESY OF:***