Swiss avalanche bulletin: automated translation with a catalogue of phrases

Kurt Winkler[1,*], Martin Bächtold[2], Stefano Gallorini[2], Ueli Niederer[1], Thomas Stucki[1], Christine Pielmeier[1], Gian Darms[1], Lukas Dürr[1], Frank Techel[1], Benjamin Zweifel[1]

[1] WSL Institute for Snow and Avalanche Research SLF, 7260 Davos Dorf, Switzerland
[2] Tele.Translator.Network, 1205 Geneva, Switzerland

ABSTRACT: In order to produce the new Swiss avalanche bulletin twice a day in four languages, a fully automated translation system was needed due to the lack of time available for manual translation. The newly developed system is based on a catalogue of predefined phrases with numerous predetermined combination options. To create an avalanche bulletin in the new system, the forecasters do not write the danger description from scratch but assemble it from a set of predefined phrases in German. Each of these phrases also exists in the database in French, Italian and English. Hence, the translations are available immediately and do not need to be proof-read or corrected. After the first operational winter season, users confirmed that the danger descriptions were correct and clear. Forecasters confirmed that it was possible to describe every situation with sufficient accuracy and within the short time span available. The cost-benefit ratio of the catalogue-based system is excellent. The savings from manual translations will exceed the initial cost of developing the phrase catalogue within a few years. A catalogue-based translation system remains limited to a very small sublanguage but is well-suited to avalanche forecasting. The database could conceivably be adapted to other multi-lingual countries or extended to other topics such as weather forecasting.

KEYWORDS: avalanche warning, automated translation, catalogue of phrases, communication

1 INTRODUCTION

A reliable avalanche bulletin must be as accurate, understandable and up-to-date as possible. Especially in the morning, the available time span between incoming field observations and the last useful editing date of an avalanche bulletin is short – too short for time-consuming manual translations. While developing the new Swiss avalanche bulletin, it became clear that an instant, fully automated translation system would be essential if we wanted to achieve our ambitious aim of issuing the bulletin in four languages (German, French, Italian and English) twice a day.

Despite efforts made around the world, the quality of commercial translation programs is still inadequate for the publication of warnings, particularly when there is no time for proof-reading or corrections and when experts for all the languages are not available in operational service.

The daily Swiss national avalanche bulletin was manually translated for many years. A comparison with the Canadian TAUM-Météo translation model (e.g. Isabelle, 1987, Lepsus et al., 2004) shows that the sentences collected over all those years are nowhere near comprehensive enough to be used directly as a catalogue of phrases for describing current avalanche dangers. Consequently, a custom-made, fully automated real-time translation system was developed based on a catalogue of standard phrases. Since November 2012 it has been used by the Swiss avalanche service to translate the danger descriptions in the avalanche bulletins from German into French, Italian and English. Catalogue-based translation systems have been used before (e.g. Schug, 2010), but mostly with smaller and less complex data sets.

In chapter 2 of this paper, we touch on some of the specific features of the Swiss avalanche bulletin. We then focus on the creation and use of the newly developed translation system. As languages are extremely complex, it is generally possible to translate segmented phrases but we could not find any theoretic or scientific approach for developing a whole catalogue of phrases. Chapter 3 describes how we created our catalogue in an empirical way, from the construction of the German sentences to the translations themselves and the various quality tests.

With neither a theoretical background nor practical examples to rely on, and with only a year of preliminary investigations, there was no guarantee at the start of the project that a catalogue of phrases would actually work in operational avalanche warning. In chapter 4, we take stock of the situation after the first winter of

Corresponding author address:
Kurt Winkler, SLF, Flüelastrasse 11,
7260 Davos Dorf, Switzerland
Tel: +41 81 417 01 27
E-mail: winkler@slf.ch

operational use: the whole system works better than we dared to hope. In chapter 5, we summarise the most important points to consider when constructing a catalogue of phrases and discuss the possibility of adapting this system to other countries and topics.

2 THE NEW SWISS AVALANCHE BULLETIN

The Swiss avalanche bulletin underwent a complete overhaul at the beginning of winter 2012/13. The key changes were improved visualisation with clear graphics and danger descriptions published twice a day in four languages (Figure 1).

The new structure strongly reflects the 'information pyramid' as recommended by the European Avalanche Warning Services (EAWS, 2009). Users can see the most important information at a glance: first the danger level, then the elevation zones and aspects where the danger is particularly prominent, followed by the danger descriptions (Figure 2). Up to this level, the content is updated twice a day and automatically translated using the catalogue of phrases. The next level of the information pyramid, namely the description of snow cover and weather, is issued only in the evening and is translated manually. Automatically measured snow and weather data are at the lowest level of the information pyramid. They are available at any time but are not checked by the avalanche service.
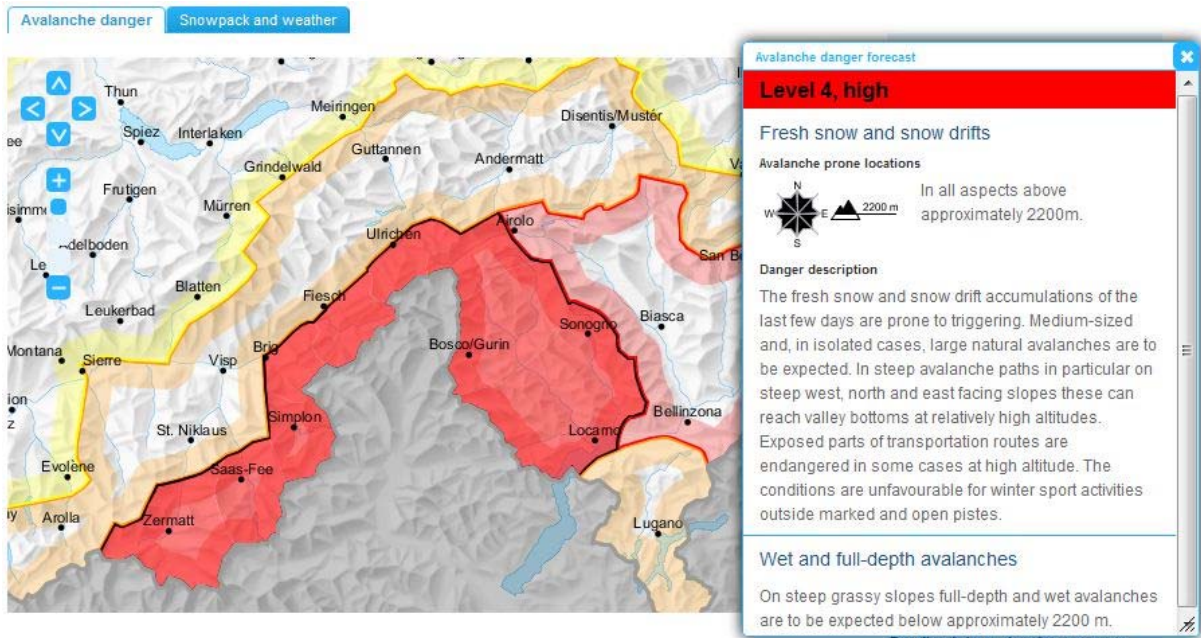

Figure 1: Detail from the Swiss avalanche bulletin with the danger description for the highlighted region. The text originates from the catalogue of phrases (here in English). More examples in all languages are available at www.slf.ch (in summertime only as .pdf in the archive: www.slf.ch/schneeinfo/Archiv)
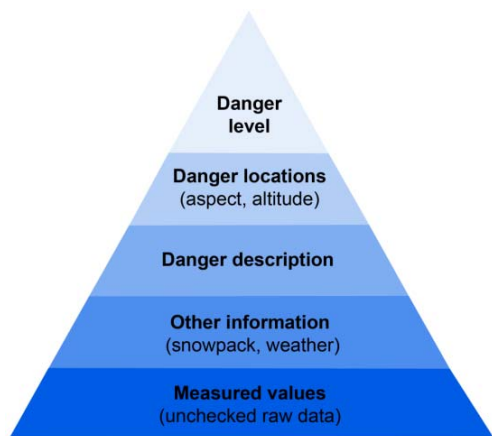

Figure 2: The most important topic first. Information pyramid as recommended by EAWS.

3 CATALOGUE-BASED TRANSLATION SYSTEM

A catalogue-based translation system is a collection of predefined phrases and cannot therefore be used to translate arbitrary sentences. The phrases were created in the source language, translated manually into one or more target languages and stored in a database. To generate a text, the users do not write it from scratch but assemble it from the set of predefined phrases in the source language. Once a phrase is chosen, it is immediately available in the target languages.

3.1 Catalogue of phrases for the danger description in the Swiss avalanche bulletin

In our system, German is the source language and French, Italian and English are the target languages. The individual sentences are not static but consist of a succession of up to 10 segments. In each segment of the phrase, the forecasters can select from a pull-down menu of predetermined options for a segment. These options can likewise consist of a series of sub-segments with selectable options, and as part of the sub-segments even sub-sub-segments are possible (Figure 3). Theoretically, the 110 predefined phrases could be used to generate several trillion different sentences.

3.2 Creating the phrases in the source language

The sentences were created by an experienced avalanche forecaster whose native language is German and who has a good knowledge of all the target languages. Numerous avalanche bulletins from the past 15 years were consulted in order to cover as many situations as possible. No phrases were taken directly: their content was always generalised and the phrase structure was simplified wherever possible. The challenge was to find sentences that were universal enough to describe all the possible danger situations and simple enough to be translated. All the sentences had to be easy to understand; literary value was not a priority. As a sentence can only be used when it works in the source language as well as all target languages, the original German sentences had to obey the following rules:

- Adjectives can only be used when they refer to subjects with the same gender and number in all the options of a language. Therefore, the gender and number of the subjects must already be known in all the target languages in order to define the sentence in the source language. To simplify the catalogue, some terms were changed, e.g. the Italian word for 'full-depth avalanche' was changed from the masculine 'scivolamento da reptazione' to 'valanga da reptazione' in order to make it feminine like all the other types of avalanches and snow slides.

- Articles alter according to number and – in most of the languages used – gender and must therefore usually be included in the same option as the noun.

- Prepositions often change with the noun and must therefore also be included in the same option as the noun, e.g. 'in' Ticino (a Swiss region), but 'on the' Corvatsch (a mountain).

- As German has four cases, this sometimes necessitated splitting certain phrases into additional segments and sub-segments. In the target languages, however, the meaning is indicated through the word order or use of prepositions, so this problem does not occur.

3.3 Testing phrases in the source language

Re-writing old avalanche bulletins using the new catalogue system and generating real and imaginary situations showed that the catalogue also allows 'new' situations to be described. Special attention was paid to the description of danger levels 4 and 5 because they are rare in the previously existing dataset.

3.4 Translation of the catalogue

Clearly, the trillions of sentences that are theoretically possible cannot be translated individually. Consequently, the translations must

Segment 1	Segment 2	Segment 3	Segment 4
In all aspects		snow drift accumulations	will form.
Adjacent to the ridge line	small		have formed.
During the night	hard		
As a consequence of the strong wind	further		
In particular _sub-segment 1-5_	easily released		
at high altitude	clearly visible		
in high Alpine regions			
above the tree line			

Figure 3: Phrase schema with four segments and one sub-segment (in cyan). In this example, [blank] is one of the options in the second segment.

also contain sentences with a segmented structure. Although German, French, Italian and English are all Indo-European languages, the differences in word order, gender, declension and so on generally make segmented translation extremely difficult. Fortunately, it came to light that it was possible to split up the carefully created phrases (all of which followed the aforementioned rules) into up to 10 parts and then reassemble these segments in an idiomatic way for all languages. However, due to the complexity of the translation, specific editing and visualisation software had to be developed by the translation agency.

In addition to the omnipresent problem of inflection, ensuring the correct word order also proved difficult. Other problems to be addressed included:

- apostrophes and elisions to avoid hiatus, which are widely used in French and Italian;
- the Italian impure 's';
- the split negation in French (ne ... pas).

When translating the individual sentences and options, no logical functions, distinction of cases or post processing were used, except for a check to ensure the presence of a space between the different segments and a capital letter at the beginning of each sentence. In comparison with the source language, only two changes were allowed in the target languages (Figures 4 and 5): the segment order could vary between the languages and each segment could be split in two (into ...a, ...b). The latter facility was widely used, mainly to construct idiomatic word orders. Unfortunately, this split restricts the use of our catalogue to translations from German into the other languages. Technically, the system could be easily used for translations in the opposite directions too, but it would be difficult for the forecasters to find the correct sentences in a source language that contains segment splits.

In addition to standard translation techniques, the following methods were used:

- replacing adjectives with participles or adverbs, as these are not affected by the number and gender of the subject;
- changing from an active to a passive formulation or vice versa;
- regrouping the words in the different sentences;
- using synonyms;
- in isolated cases, and only when not essential to the content, options were abandoned in all languages.

3.5 Testing the output of the translation catalogue

Not all possible sentences are meaningful, but all those that make sense content-wise must be correct in all languages. As no proof-reading is possible in operational use, the translations in the catalogue must be guaranteed to be of high quality, even though the trillions of possibilities cannot be checked individually. Special attention had to be paid to the optional segments, where [blank] can also be chosen. Three different quality checks were conducted for each phrase:

- A sentence generator was created to randomly generate a number of permutations from a phrase. The generated sentences were individually checked by the translators.
- All the options of every segment, sub-segment and sub-sub-segment were checked sequentially by the author of the phrase catalogue in all four languages together. This involved checking both the correctness of the phrases and the content of the translations.

Fictitious avalanche bulletins were written and checked by native-speaker avalanche experts. The importing of the dataset into the database and the newly developed bulletin-editor were also tested.

Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Mit dem Schneefall	steigt die Gefahr			an.
An steilen, wenig befahrenen Sonnenhängen		ab dem Morgen	allmählich	
Im Hochgebirge		am Mittag	rasch	
In diesen Gebieten		im Tagesverlauf	vorübergehend	

Figure 4: Schema of a simple phrase without any sub-segment in the source language (German).

Segment 3	Segment 1	Segment 2a	Segment 4a	Segment 2b	Segment 4b
	as a consequence of the snowfall	there will be	an	increase in the danger	.
From the early morning	on steep, little used sunny slopes		a gradual		.
In the middle of the day	in high Alpine regions		a rapid		.
As the day progresses	in this regions		an		for a short time.

Figure 5: The phrase of Figure 4 in English, one of our target languages: the order of the segments is different and segments 2 and 4 are split.

4 OPERATIONAL USE AND FEEDBACK

During the first winter of operational use of the catalogue of phrases with approximately 1,000 danger descriptions per language, only marginal corrections were required. Users confirmed that the danger descriptions were correct and clear.

The Swiss avalanche forecasters took stock:

- As before, the content of the danger descriptions in German was proof-read and discussed by at least two avalanche forecasters. Once the content of the German text was found to be correct, the translated text was published without any further proof-reading or corrections (Stoffel and Niederer, 2013).

- It was possible to describe every situation with sufficient accuracy. In the case of missing sentences, it was possible to add arbitrary text strings in all four languages and use them immediately. However, no such 'joker phrases' were actually used during operational service.

- From childhood onwards, people are accustomed to formulating and writing phrases. Searching for phrases in a catalogue is quite different. Therefore, the main challenge in operational service was not missing sentences but rather finding the matching sentences quickly enough.

- Nevertheless, it was possible to describe every situation within the short time span available. The search engine incorporated into the custom-made bulletin-editor proved essential here (Ruesch et al., 2013).

5 CONCLUSIONS

The newly developed catalogue-based system proved to be well-suited to automatically and instantaneously translating the danger descriptions in the Swiss avalanche bulletin from the source language (German) into the three target languages (French, Italian and English). The translations do not need to be proof-read or corrected. It was important that the developer of the structured source language phrases had a solid knowledge of the target languages.

Users and avalanche forecasters confirm that the danger descriptions are correct and clear. However, the quality of the language still needs to be assessed. This could be done by comparing warnings from the old, manually translated avalanche bulletin with warnings from the new, catalogue-based bulletin in a blind test performed by native-speaker users.

A catalogue-based translation system can be very efficient, especially where there is not enough time for manual translation. However, the construction and translation of a phrase catalogue is demanding and the developers must have experience in writing similar texts from scratch as well as knowledge of the target languages.

The crucial point during operational service is to find the matching phrases quickly enough. Choosing the correct options within a sentence proved to be easy and fast. The implemented search engine was essential. Nevertheless, a catalogue should be limited to as few phrases as possible. This requires the individual phrases to be as universal as possible and restricted to a very small sublanguage, which is the case for the specialised topic of avalanche forecasting.

As using a phrase catalogue requires experience, frequent operational use is necessary. Financially, the cost-benefit ratio of the catalogue-based system is excellent. The savings from manual translations will exceed the initial cost of developing the phrase catalogue within a few years. Applying the database to other multi-lingual countries (e.g. Canada, Italy) or extending it to topics such as weather forecasting is conceivable but would require further development.

6 REFERENCES

EAWS, 2009. Report of results. 15th European Avalanche Warning Services Conference, Innsbruck, 16-17 June 2009.

Schug J., 2010. Personal communication, Meteomedia, Gais, Switzerland, 14 May 2010.

Lepsus T., Langlais P. and Lapalme G., 2004. A corpus-based Approach to Weather Report Translation. Technical Report, University of Montréal, Canada.

Isabelle P., 1987, Machine Translation at the TAUM Group. In M. King (ed.) *Machine Translation Today: The State of the Art*, Edinburgh: Edinburgh University Press, pp. 247–277.

Stoffel A. and Niederer U., 2013. Software tools developed for the Swiss avalanche warning service - a system overview. Proceedings ISSW 2013. International Snow Science Workshop, Grenoble, F, 15-19 October 2013, this issue.

Ruesch M., Weiss G., Gerber M. and Egloff A., 2013. The software behind the interactive display of the Swiss avalanche bulletin. Proceedings ISSW 2013. International Snow Science Workshop, Grenoble, F, 15-19 October 2013, this issue.