revista.tradumàtica
tecnologies de la traducció

205-210

MT and Proper Nouns: How a German Model Became a Boat Operator

Barbara Inge Karsch
BIK Terminology

ABSTRACT

Writers and translators have difficulties treating proper nouns correctly. These designations represent concepts that are very likely not common knowledge. While humans can research, machines can only apply data provided. It is therefore important that proper nouns are documented in term bases and made available to MT engines.

Keywords: machine translation, terminology, proper nouns, individual concepts

RESUM *(La traducció automàtica i els noms propis: com un model alemany es converteix en un operador de vaixell)*

Tant redactors com traductors tenen dificultats per realitzar el tractament correcte dels noms propis. Aquestes denominacions representen conceptes que probablement no pertanyen al coneixement comú. Mentre que els humans poden recercar el concepte, les maquines només poden aplicar les dades de què disposen. Per aquest motiu, és important que els noms propis estiguin documentats a la base de dades terminològiques i que estiguin a disposició dels motors de traducció automàtica.

Paraules clau: traducció automàtica, terminologia, noms propis, conceptes individuals

RESUMEN *(La traducción automática y los nombres propios: cómo un modelo alemán se convierte en un operador de barco)*

Tanto redactores como traductores tienen dificultades para realizar el tratamiento correcto de los nombres propios. Estas denominaciones representan conceptos que probablemente no pertenezcan al conocimiento común. Mientras que los humanos pueden investigar el concepto, las máquinas únicamente pueden aplicar los datos de los que disponen. Por este motivo, es importante que los nombres propios estén documentados en una base de datos terminológicos y que estén a disposición de los motores de traducción automática.

Palabras clave: traducción automática, terminología, nombres propios, conceptos individuales

1. Introduction

In 1998, when J.D. Edwards had just switched to e-mail as the main electronic communication system, when we started doing terminological research online, and when the first machine translation applications were available on the internet, a few friends and I had fun with MT. I would type an e-mail in German, my native language, send it to a few American colleagues and see what they would get from that. I am sure most of us have played this type of Chinese whispers via MT at some point in our careers. It was very entertaining back then, and while MT systems have improved in many aspects, it can still be entertaining today.

Mind you the messages back and forth were not always work-related. And one instance still comes to mind today: The name of German super model, Claudia Schiffer, was translated as Claudia boat operator. It is obvious that a mistake happened. The MT engine, I believe we used Babel fish at the time, did not recognize that the words Claudia and Schiffer form a lexical unit, let alone a name. Most of us would have recognized Claudia Schiffer as the name of an individual. But unless you were involved in the creation and naming of a new concept, it is not that obvious even to humans that a particular combination of words is the name of a new product or device. And it is even more difficult, if the words represent a string on a software screen.

For terminologists or those in a terminologist role, the distinction between an individual concept and a generic concept is harder to make than we might think at times. Very likely translators can relate to it because they have to analyze the concept behind the terms and names in a text very quickly. It is also clear that those designing products, consequently naming concepts and/or writing about them, often don’t know the difference. And it would be very hard for those doing post-editing on machine translation to fix problems that concept creators, writers, and the MT engine even with support by terminologists didn’t get.

In this short paper, we will analyze why Claudia Schiffer turned into a boat operator by defining the main concepts and relating these concepts to one another. We will look at a small text sample and its MT translation specifically with the focus on proper nouns. In conclusion, suggestions for those at the cross-roads of terminology management, MT and post-editing will be made. The perspective will be that of a text that needs to be translated from English as a source language into other languages.

2. Definitions

The official ISO definition for terminology work is “work concerned with the systematic collection, description, processing and presentation of concepts and their designations” (ISO 1087, 2000:10). To establish terminology for a project, we collect concepts and terms, research and document them in a terminology management system (TMS), and then distribute and use the data in many different applications. One of these applications could be an MT system. It is important to note that at the center of terminology work is the concept, which we have to understand before we can name it in a source and/or target language.

A concept is “a unit of knowledge created by a unique combination of characteristics” (ISO 1087, 2000:2). A concept that corresponds to one object in the world is called an individual concept (ISO 1087, 2000:2). An example would be Windows 8. There is only one operating system in the world known as Windows 8. A general concept, on the other hand, is a concept “which corresponds to two or more objects which form a group by reason of common properties” (ISO 1087, 2002:3). An example of a general concept is the one represented by the term “operating system.” There are multiple types of operating systems (e.g. Linux, Microsoft Windows, Mac OS, and UNIX) and they all are defined as collections of software which provide services to programs and manage hardware resources. Each one might do more than that, but that is roughly what they all do.

Machine translation can be defined as a process in which natural language content in a source language is translated into content in a target language using computers and without human intervention. The focus point here is the transfer of content. MT engines transfer content from one language to another by looking at text segments and their words and terms in a source text and then matching them to words and terms or segments used in a target language. The process does not involve the detour, if you will, via the concept. In other words, MT stays on the linguistic level.

On that linguistic level, we distinguish various types of designations. A designation is a representation of a concept by a sign which denotes it (ISO 1087:2000, 6). In other words, the concepts discussed in a text are represented by a designation. There are two types of designations that we should look at, namely terms and appellations.

A term stands for a general concept, e.g. operating system is a technical term. Windows, on the other hand, is an appellation, more commonly referred to as name, of an individual concept. A term or an appellation has a part of speech. In terminology, we mostly deal with nouns, verbs and adjectives, as they are the main carriers of information and represent concepts. 80% of the terminology managed in most databases is nouns. Nouns can be divided into common nouns and proper nouns where common nouns stand for general concepts and proper nouns stand for individual concepts.

Proper nouns are often capitalized in English, but casing is not a reliable indication that something is a proper noun. Often times, proper nouns are not translated, and a translator or a machine translation engine would need to know that that is the case. In their analysis, source terminologists understand the concept behind a designation and indicate in the database whether something is a general or individual concept and more explicitly whether something is a common or proper noun. They may even indicate that a name is trademarked. Target terminologists use that information to research the concept in the target language and identify the correct designation, again either a proper or common noun.

3. Application

Translators use their implicit knowledge of the subject matter to identify general and individual concepts. They will also research extensively those concepts that are new in a subject area or new to them. Or they may use the terminology database prepared for them. One of the most common mistakes in databases and in translated material is that proper nouns in the source language referring to individual concepts, e.g. a product names, are “translated” by common nouns. This mistake is indication that, for example, someone new to the terminologist role did not analyze the concept sufficiently or even understand the concept. It may also indicate that the difference between proper and common noun and what they signify isn’t clear to them.

If this is difficult for humans, how does an MT engine deal with it? Of course, it depends on the system. Since machines do not possess knowledge or do research, they must rely on the terminology prepared for them. That can either be reflected in the translation memory or in a TMS. If the humans that prepared the TM or the term base understood it, the chance that the MT engine will get it right is high.

The following is an example of an English source text that was translated with a variety of MT engines into German. The proper nouns in the source as well as in the target texts are underlined for easy identification. The translation listed under 3.5. was supported by a list of proper nouns in the source and target languages.

3.1 Source text (Source: http://support.microsoft.com/kb/2745101)

Assume that you debug a Windows Store app that supports the share target contract in Microsoft Visual Studio 2012. In this situation, you receive an error message that resembles the following when you try to use the Share charm in Windows 8: Something went wrong with Share. Try again later.

3.2 German translation from Google translate, Oct 14, 2012

Davon ausgehen, dass Sie eine Windows Store App, die Aktien Bausparsumme in Microsoft Visual Studio 2012 unterstützt debuggen. In diesem Fall erhalten Sie eine Fehlermeldung, die der folgenden ähnelt, wenn Sie den Anteil Charme in Windows 8 zu verwenden: Etwas ist schiefgelaufen mit Share. Versuchen Sie es später erneut.

3.3 German translation from Bing Translator, October 14, 2012

Genommen Sie an, Sie Debuggen eine Windows-Speicher-app, die Anteil Zielvertrag in Microsoft Visual Studio 2012 unterstützt. In diesem Fall erhalten Sie eine Fehlermeldung, die die folgenden, ähnelt Wenn Sie versuchen, den Share-Charme in Windows 8 zu verwenden: Etwas schiefgegangen mit Anteil. Versuchen Sie es später erneut.

3.4 German translation from Systran, without terminology

Nehmen Sie an, dass Sie eine Windows-Speicher-APP ausprüfen, die den Anteilzielvertrag in Microsoft Visual Studio 2012 unterstützt. In dieser Situation empfangen Sie eine Fehlermeldung, die das folgende ähnelt, wenn Sie versuchen, den Anteilcharme in Windows 8 zu verwenden: Etwas ging mit Anteil schief. Versuchung noch einmal später.

3.5 German translation from Systran supported by a list of the correct proper nouns in source and target languages

Nehmen Sie an, dass Sie eine Windows Store-APP ausprüfen, die den Anteilzielvertrag im Microsoft Visual Studio 2012 stützt. In dieser Situation empfangen Sie eine Fehlermeldung, die das folgende ähnelt, wenn Sie versuchen, den Charm Teilen in Windows 8 zu benutzen: Etwas ging mit Charm Teilen schief. Versuchung noch einmal später.

The following table gives an overview of the source terms from the text and how the individual systems dealt with it. The right-most column lists the target names provided to the Systran engine for the translation that resulted in example 3.5.

Source name	Google	Bing	Systran	Systran plus terminology	Correct target name
Windows Store	Windows Store	Windows-Speicher	Windows-Speicher	Windows Store	Windows Store
Windows Store app	Windows Store App	Windows-Speicher-app	Windows-Speicher-APP	Windows Store-APP	Windows Store-App
Microsoft Visual Studio 2012	Microsoft Visual Studio 2012	Microsoft Visual Studio 2012	Microsoft Visual Studio 2012	Microsoft Visual Studio 2012	Microsoft Visual Studio 2012
Share charm	Anteil Charme	Share-Charme	Anteilcharme	Charm Teilen	Charm "Teilen"
Windows 8	Windows 8	Windows 8	Windows 8	Windows 8	Windows 8
Share	Share	Anteil	Anteil	Charm Teilen	Charm "Teilen"

4. Evaluation

None of the systems had any problem with the proper nouns Microsoft Visual Studio 2012 and Windows 8. Very likely, both basic product names have been around long enough to be part of translation memories in any of the systems. The only additions, i.e. the versions 2012 and 8, were easily performed by all of the systems.

It is more interesting to examine the name Windows Store. Bing and Systran recognized Windows as the proper name, but translated Store literally; although both used hyphens to connect the name Windows and the word Store, which indicates that the engines had information about a relationship between the two. In both cases, hyphens were added to connect the name to the term app as well.

While Windows Store is a proper name the combination Windows Store app is not. It simply represents a type of application; there are many objects that are captured by this general concept. It is not a surprise that the Systran version with terminology did not get the full term right. After all, for that run, the German equivalent of the term app, which would be App, was not provided explicitly. The correct German term for the concept designate in English by the source term Windows Store app is Windows Store-App.

Another interesting case that is similar, yet not the same is the proper name Share. Share is a type of charm, since this concept only exists once, it is an individual concept. Often times, translators ask whether xyz (here “charm”) is part of the name. While there is probably no rule for this, understanding in the target language is very likely enhanced, if xyz is treated as part of the name. In this example, if the reader doesn’t know what Share is, understanding is diminished.

At the root of this last example is the fact that the source sentence “Something went wrong with Share.” is very ambiguous and clarification would be needed in the source text. Share in this sentence could refer to the Share charm, although it is hard to imagine that “something goes wrong” with a static thing such as a charm, which is a sort of button on the interface. Rather it seems that the author refered to the share process triggered when the user clicked the Share charm. This kind of confusion where a name of an individual concept is also used to refer to a process that might be triggered by a concept by the same name is very common in software-related content in English.

5. Conclusion

From this short example we can see that an understanding of the difference between individual concepts and general concepts is important. It would allow writers to be clearer in their treatment of proper nouns, but also of common nouns in the source text. Once that is given, translators and terminologists could clearly identify these concepts and avoid faulty entries in a terminology database or incorrect literal translation in a text. When humans are clear about the distinction, translation memories are more reliable and term bases can be more complete.

For rules-based MT engines it would be beneficial to set up an individual concept with its appellation in the MT dictionary correctly. Once set up, it would allow the engine to apply it consistently throughout the text. For statistical MT engines the benefit would come not only from the terminology resources added in the system and given a higher weight, but also from cleaner TMs provided by human translators.

Even if clean TMs and terminology were used in a project, focusing on proper nouns and individual concepts during the post-editing process may be worthwhile. Of course, the difference must be clear to post-editors. But they must also understand that the author may not have been clear on it and may have refered to concept ambiguously in the source text.

Household names have long since been part of the vast volume of material stored in publicly available MT engines, such as Bing Translator or Google Translate. So, there is no more fun to be had even with names, such as the one of Brazilian super model, Gisele Bündchen: Literally translated from German into English, her name would be Gisele small cuff. But it is worthwhile looking at proper nouns of products in MT for technical texts no matter what system is used: Even a Windows product was new at some point in time, and it would be extremely embarrassing to have it translated literally.

6. References

International Organization for Standardization. "International Standard 1087-1." Terminology work - Vocabulary - Part 1: Theory and application. Geneva, 2000. Vol. ISO 1087-1:2000(E/F).