Do you know English? The challenge of the English language for patent searchers

Stephen van Dulken

Flat 5, 17 Claremont Gardens, Surbiton, Surrey KT6 4TL, UK

ARTICLE INFO

ABSTRACT

The complex and inconsistent nature of the English language presents problems for patent searchers researching the prior art. This is true for native speakers as well as for those who use it as a second language. These problems include confusion in translations; "Patentese", the jargon used by patent attorneys; terminology, which can take time to be adopted; "faux amis", words which you think you know as they look identical to foreign words; the oddities of English spelling; multiple meanings for the same words; words that have opposite meanings; synonyms; Americanisms as different spellings and different words; words that are both nouns and verbs; compound nouns, which are often spelt as two words; spelling mistakes; and syntax. Conclusions suggest using broad classes together with keywords; looking for synonyms; allowing for two words in compound nouns; using adjacency operators; combining sets of results; and using citation searching as an additional search, especially if little is found, or the invention is difficult to describe. A thesaurus of recommended words and spellings would be useful if adopted by those preparing abstracts.

© 2014 Published by Elsevier Ltd.

1. Introduction

This is a revised, and somewhat expanded, version of a presentation I made at the European Patent Office's *Patent Matters 2014*, a conference that provides training in patent searching, in April 2014. It was titled *Do you know English? The challenge of the English language for searchers.*

It is a personal view of the linguistic problems that English (and indeed to a large degree any language) presents for patent searchers looking through the prior art. This is particularly true for those who use English as a second language, but native speakers also frequently have problems. It is a complex, rich and inconsistent language, part Germanic and part Latin in origin, with a huge vocabulary and numerous synonyms.

2. My background

My own background inevitably influences my attitude to search problems and how to solve them. I am a native English speaker, educated in a British curriculum (and hence its dialect), but I lived in the USA between the ages of 10 and 15. Hence as a patent searcher I instinctively used, or looked for, American variants in wording or spelling. I have no background in linguistics.

E-mail address: svandulken@hotmail.com.

I am now retired from full-time work but worked at the British Library as a patent specialist from 1987 to 2013, for the last ten years carrying out numerous priced patent searches in the mechanical and electrical areas. My academic background was as a historian: not ideal, I freely admit.

These were mainly prior art searches for private inventors or small companies. Sadly, many clients were somewhat naïve, and did not understand the nature of prior art or its significance. Many assumed a quick search was all that was needed before investing a great deal of time or money in an invention (or asking others to do so).

The databases I used were the priced databases Derwent WPI and TotalPatent. Normally searches involved initially using Espacenet for "quick and dirty" searches (plus citation searching if only a few patents seemed to be relevant). Then a complex search strategy using classification and keywords was used on, initially, Derwent, but in the last couple of years mainly the TotalPatent database. I was supposed to spend just an hour, but it was rare for me to take less than 2 h as I was never happy with a quick search. Negative results do not necessarily mean it's not there, sometimes the wrong approach is being used.

Like other searchers I used patent classification to limit the number of hits and to avoid too many false drops. However, I found that the relevant material, especially in electronics, was often very widely scattered. The broader, industry-based classes used in

Derwent WPI were a big help here. I normally combined a set of results from a few narrowly defined classes together with loosely relevant wording with a set of results using broader classes with tightly defined wording. I frequently used forward citation searching of highly relevant results to provide additional hits.

In my final years at work, I found the ECLA (now Cooperative Patent Classification) schedules very helpful. Searching for either mobile phone or cell phone find the same relevant class area, which is obviously ideal.

3. Literature review

There is relatively little literature on the subject of language problems in patent searching

Wittman [1] gave an interesting example in Sicherheits(ski) binding, where the translation is "safety(ski)binding". He found 148 hits, but when looking for the relevant "release binding" found another 52. There were other German to English problems, as some words do not have exact equivalents in other languages. He gave the example of "gezwirnter Faden", roughly twisted thread, but not exactly. He said that some German words in patents are made up, combining three terms! He wanted a German text database (he now has it with Depatisnet).

Lyon [2] and Hull [3] have also written on the subject.

Nijhof [4] analysed what any good patent searcher does instinctively. There are three types of search terms, Type 1 are distinctive enough to be used on their own. Type 2 are usable if combined with a second term to form a distinctive concept. Type 3 are usable with two additional concepts. By analysing the possible search terms into groups, you will not mix say a Type 2 word with a Type 3 word.

He also discussed how important it is to gather together synonyms and near synonyms, including often some very odd sounding wording. The problem I found with this in my own searching was that again and again although my recall of relevant data indeed improved as I added in every more terms, the number of hits massively increased. That is, precision was much worse. The searcher and client (if different) must decide on the right balance between these dilemmas of missing relevant data and being swamped by rubbish. My own preference tended to be adding oddly worded patents, often found in citation searching, to the final hit list rather than relying on obscure wording to gather them into a search.

Adams [5] and [6] also discussed precision and recall as problems. In [6] he pointed out that in German the plural noun Fliegen, flies, is distinguished from the verb fliegen, to fly, by the use of an upper case letter for nouns. However, databases invariably do not make this distinction, and allow for either upper or lower case usage by searchers. Hence that potential help is lost when searching in German text. He added that trade names such as Karate® can complicate the issue. My own thought on this is that an option in German language databases could be looking for the first letter of a word (other than the first in a sentence) as upper case to restrict hits to nouns. I have often wondered if something similar could be done in English patent abstracts with coding being applied by an editor.

Adams also commented briefly on the problems with handling accents and the like. German words with umlauts, for example, are usually handled by inserting an extra letter in English text. If you are not aware of this you won't find it.

4. Faux amis

Faux amis are words that you think you know in foreign languages, but which mean something else. They are a problem when searching, or analysing results, in a foreign language, or if composing a search in a foreign language.

In French, for example, demander does not mean to demand, it means to request; crayon is a pencil and not a crayon; raisin is a grape and not a raisin (yet pain aux raisins has raisins, not grapes); and cortège does not mean, in English, processions but exclusively *funeral* processions. The verb commander is not used in French for "to command" but as in for example placing an order for food or drink. In English order of course has, besides the connotation of placing items or concepts in a specific arrangement, the meaning of either *placing* an order (requesting it) or the harsher *giving* an order (when in charge of staff). Subtle distinctions such as these must be confusing for those with English as a second language.

It is an urban legend that President George Bush Jr. once said "The problem with the French is that they have no word for entrepreneur." But it is true that they have no word for scientist, and make do with savant, which means a learned person. And which, in English, is reserved for an eccentric genius.

As so often we rely on context, but sometimes a search will give us a false drop as the context is only apparent with the results, not in the search terms. That is why using additional terms, or classification, is so helpful to reduce the number of false drops.

5. Machine translations

Other examples of *faux amis* occur when attempts are made to translate material. Confusion can occur and may not be apparent until queries are made. Here are a couple of examples, one where confusion occurred, the other where confusion was averted.

When the Italian astronomer Schiaparelli announced in an 1877 academic paper that he had seen canali on the planet Mars, English speakers for decades thought that he had seen canals rather than the actual channels that the Italian meant. An error had been made in the translation into English.

When the film *King Kong* was shown in Denmark, it was called *Kong King*. This was necessary, because "Kong" means "King" in Danish.

Foreign language material can be searched for using English language abstracts or by using machine translations. I remember in 1988 scoffing when someone from the European Commission who was involved in a machine translation project explained that they were making good progress in designing syntax rules and so on. I am happy to admit that I was wrong, and that they are valuable. There were jokes at the time, of course.

For example, the saying *The spirit is willing, but the flesh is weak* became, when translated into Russian and back again, *The vodka is great but the meat is lousy*. The statement *Voici l'Anglais avec son sangfroid habituel* has the meaning of "Here is the Englishman with his usual cool air", but could literally be rendered as "Here comes the Englishman with his usual bloody cold".

More likely to be a problem in patents, it has been pointed out, would be *hydraulic rams* becoming *water sheep*.

Nowadays Espacenet's PatentTranslate, for example, is very good for translating wording in patent specifications. Although I do remember a German patent where the two key terms were not translated, but all the other words were. More seriously, some terms are not exact equivalents of each other. The normal word for species has subtly different meanings in English, French and German for example.

For search purposes I used the ability to translate wording in the priced subscription database TotalPatent. It would translate the search query into numerous foreign languages, search full-text, and translate relevant results back into English to present as results. I admit that I still feel safer with material originally published in English.