

Carabao and Text Understanding

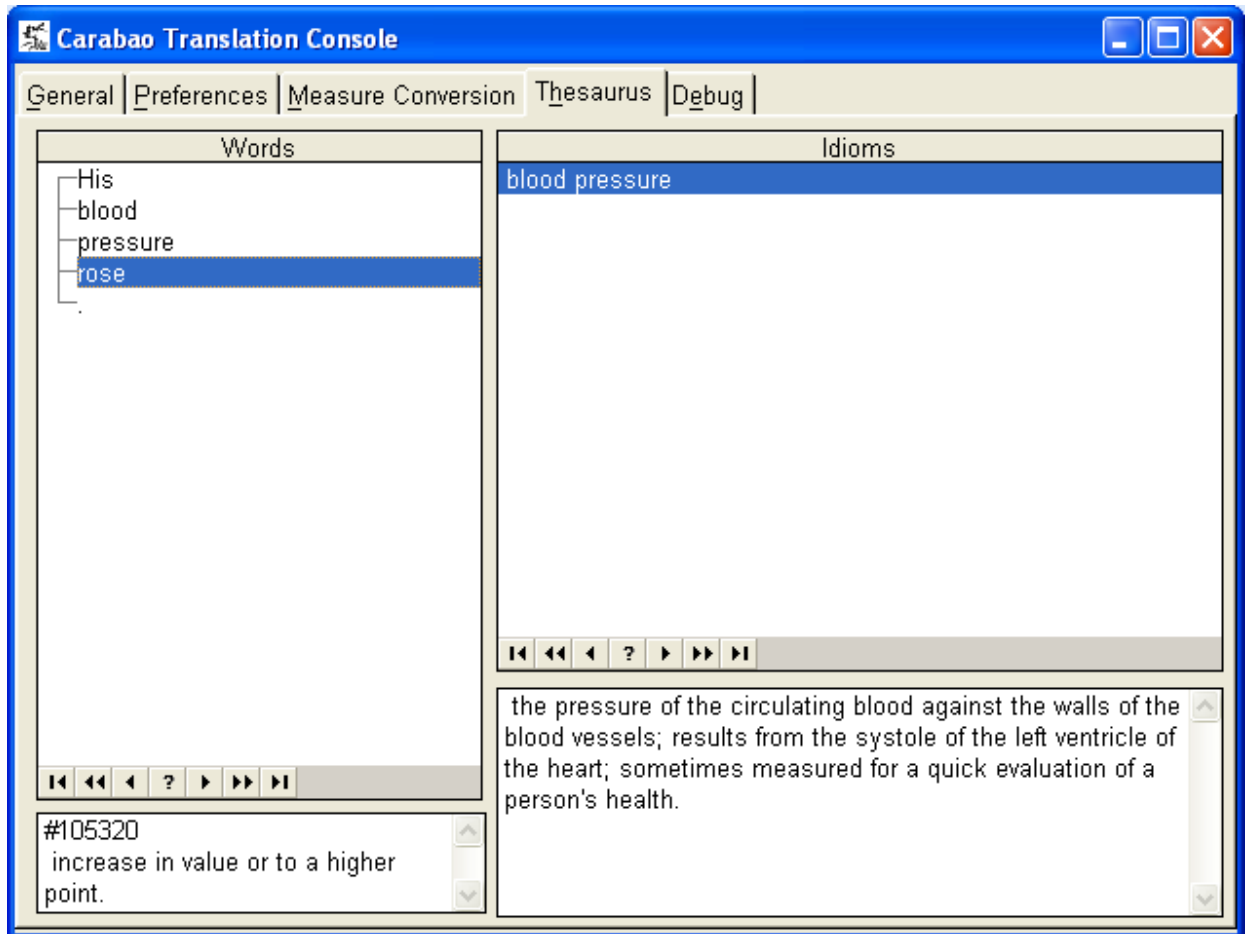
By Vadim Berman



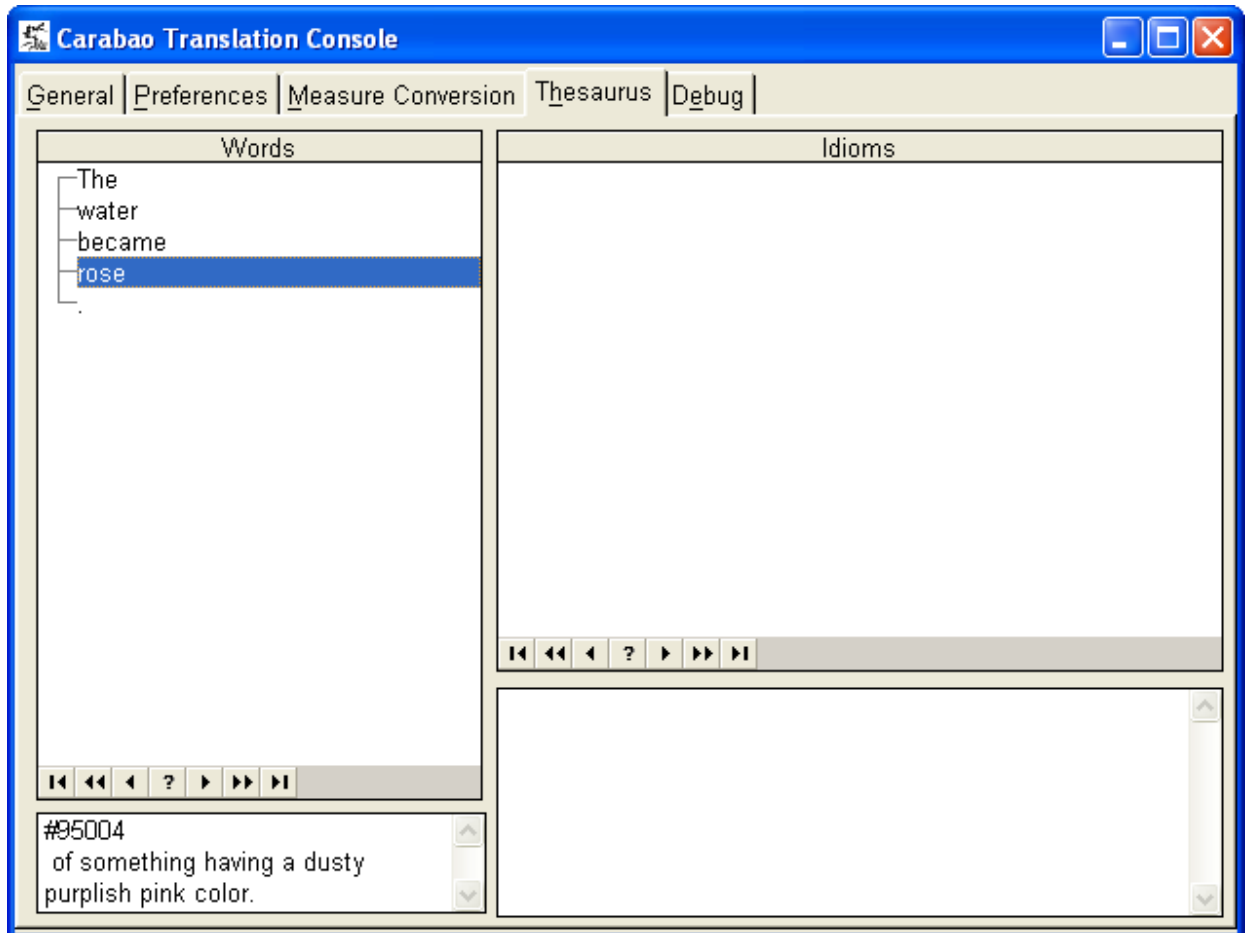
It is no secret that understanding the input text is a better way to perform common NLP tasks such as domain extraction, entity recognition or machine translation. The software methods of text understanding are very different from those employed by humans, and therefore computers face different type of difficulties. While a human being finds long, specialized terms (for instance, "polyurethane") difficult, and short ones (for example, "cut") easy to grasp - computers have problems with more ambiguous terms. And disambiguation, or choosing the correct sense of a word, is the main issue in the natural language applications.

Modern applications can do better than pick a random sense or the most frequently encountered sense. Carabao in its attempt to imitate processes occurring in the human mind employs a combination of several techniques (grammatical, statistical, and neural network approach), configurable for every language or even end-user's specific need. In simple words, Carabao plays what-ifs with grammar, semantics and current context trying to find out the exact meaning of the sentence. No matter how many features and buzzwords a language software package has to offer, the accuracy is always its most important aspect. So how good Carabao is when it comes to interpreting text?

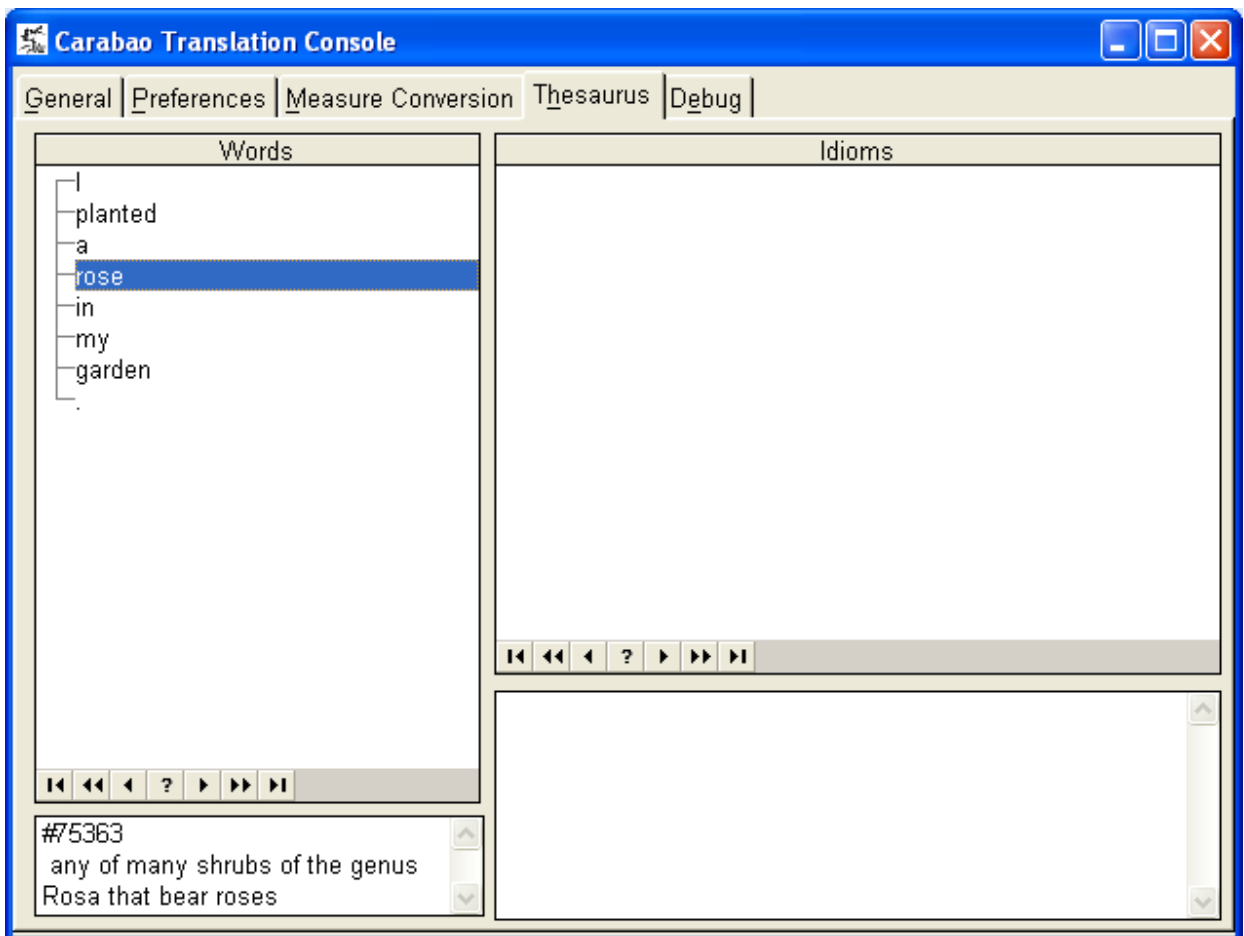
Let us examine a sentence, "**His blood pressure rose**". The word *rose* is a fine example of high ambiguity. It can be a verb, an adjective, or a noun. In every one of these cases, there are a handful of interpretations - it can be a flower, a pinkish wine, a color, etc. Carabao has thesaurus articles for most of the entries, so we can check what sense it picked (in other cases, we can use a reference number). In our case, the syntactic structure of the sentence is fairly straightforward: *rose* seems to be a verb. And this is the conclusion that Carabao has arrived at (note that the concept of 'blood pressure' is also recognized):



Let us try a sentence where 'rose' is an adjective: "The water became of rose color." Carabao is capable of determining the correct sense yet again:

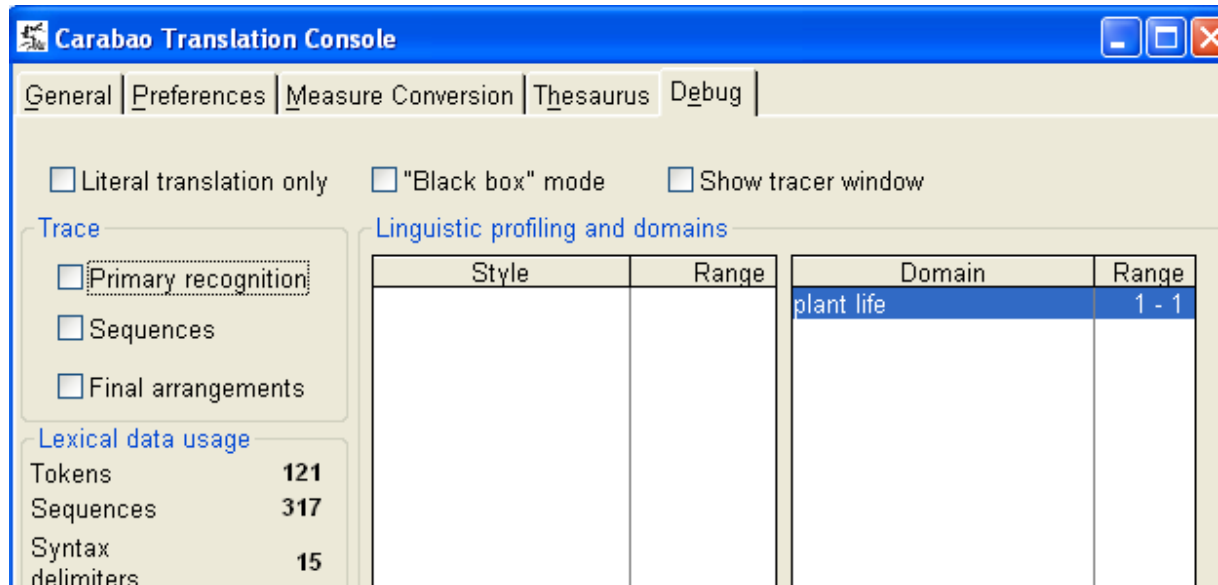


But the real challenge will be to distinguish between various meanings of the same part of speech. Let us start with something straightforward, like, "**I planted a rose in my garden**":

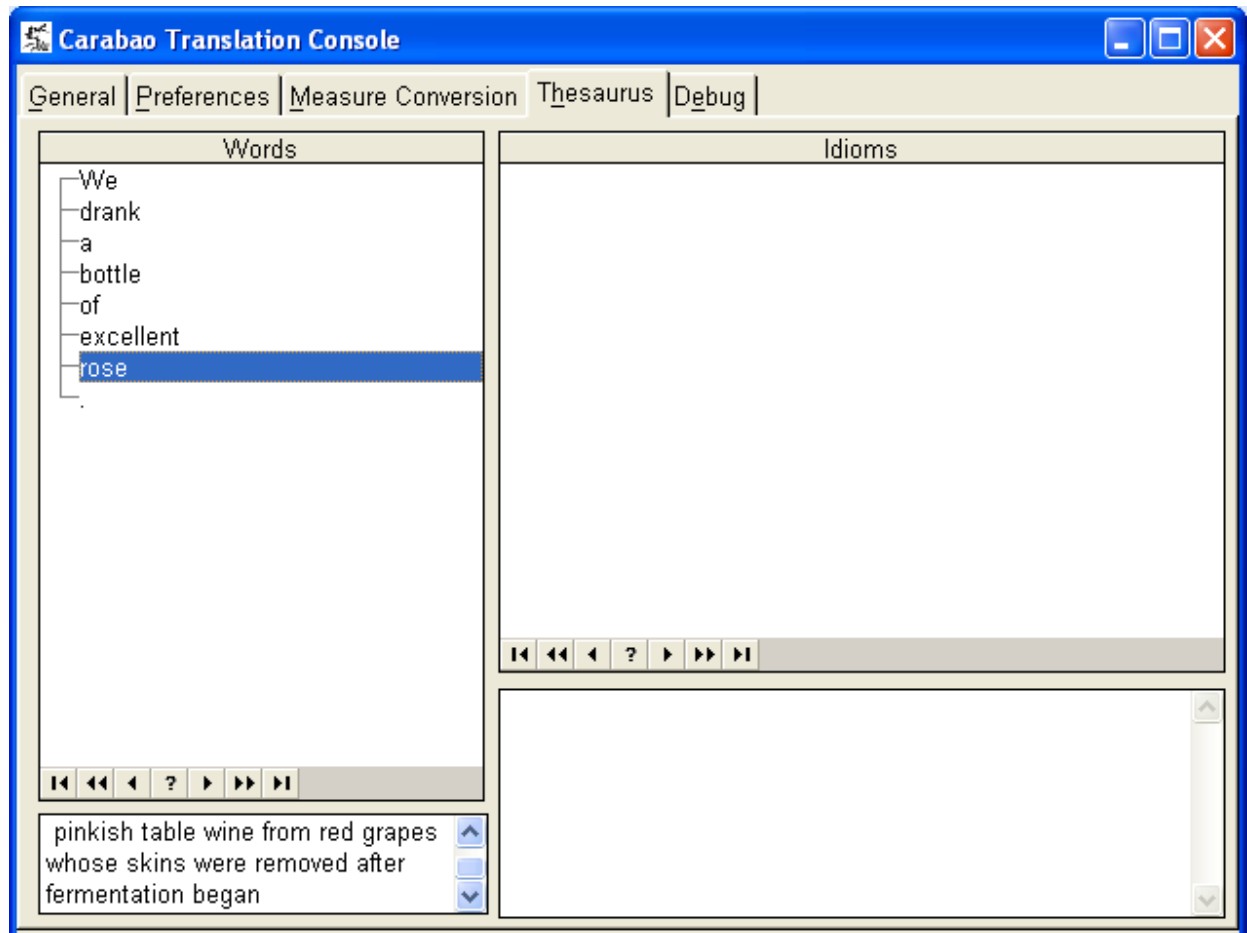


Interesting that Carabao was pedantic enough to conclude that a rose is a shrub and not a flower; which seems to be correct in our case. Here, we can highlight another feature of Carabao - extraction of domains.

Carabao is able to present domains of discourse throughout the text being processed. In this case, one sentence was enough to bring Carabao to this conclusion:



Finally, as promised earlier, let's try another sense - the rose as a kind of wine. **"We drank a bottle of excellent rose."** Carabao seems to know a lot about wine-making:



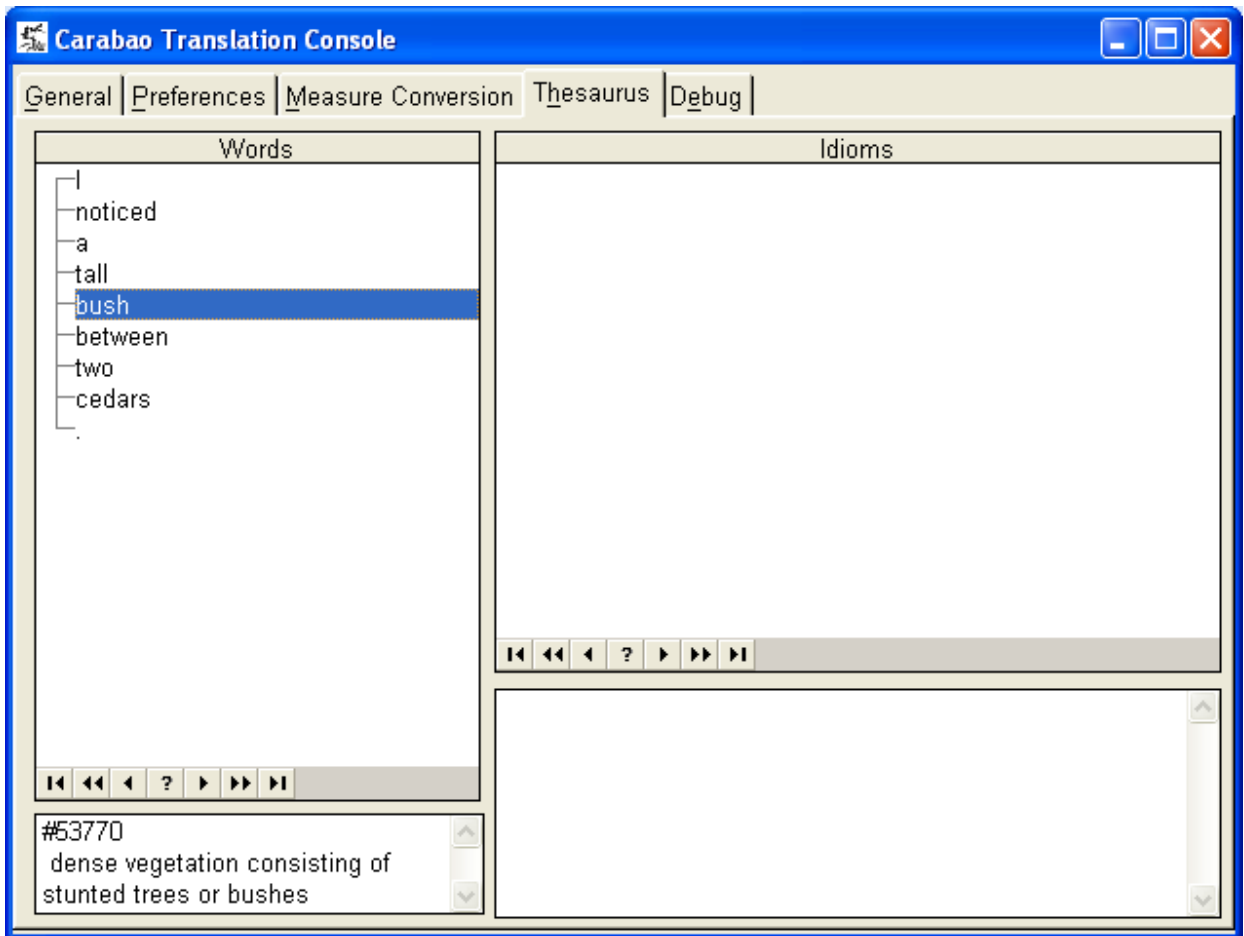
Carabao is also able to pinpoint the domain of discourse:

"Black box" mode Show tracer window

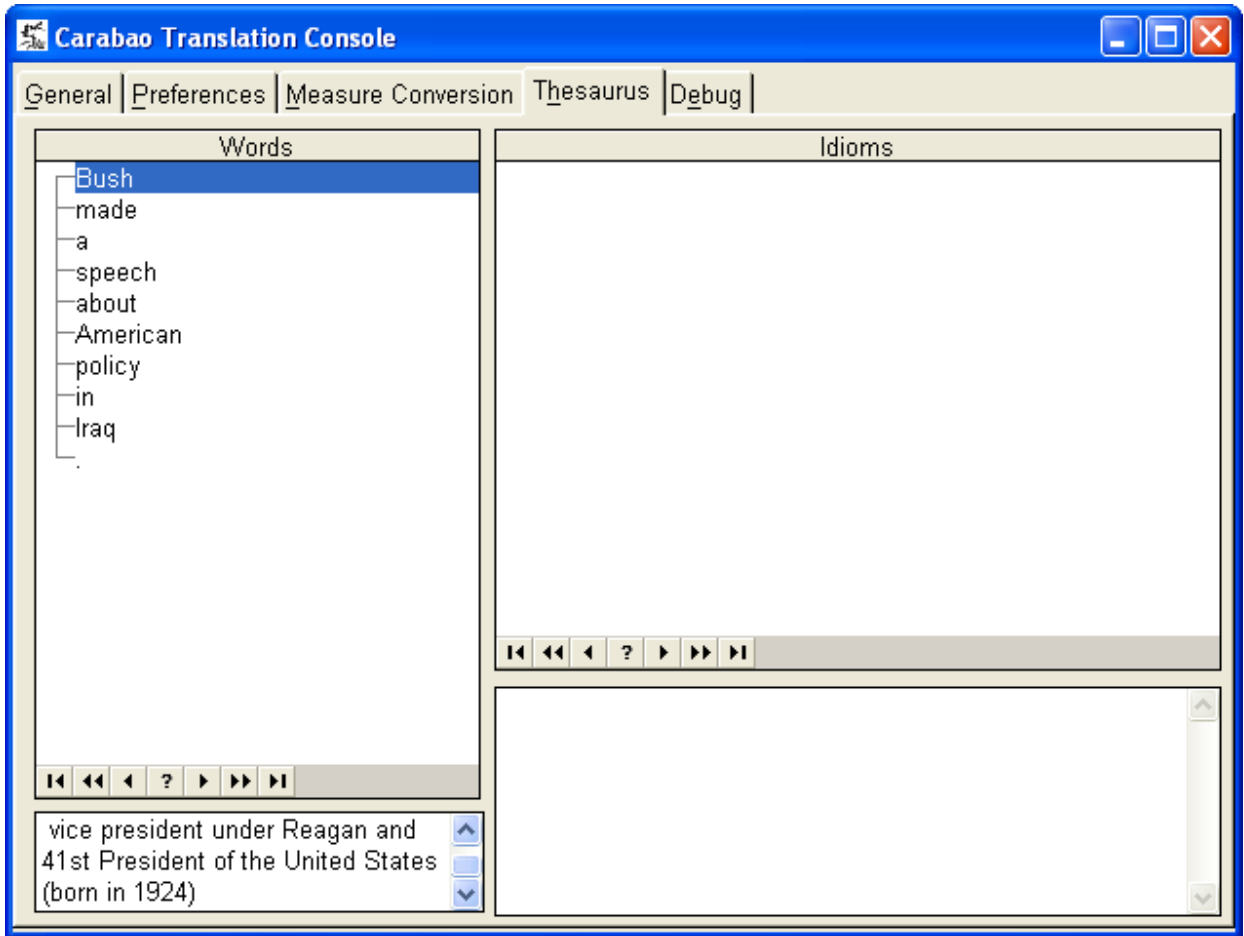
Linguistic profiling and domains

Style	Range	Domain	Range
		beverage	1 - 1

Our next experiment will be trying to analyze proper names resembling a regular noun. For example, "bush". Carabao defines a regular bush as 'dense vegetation' - for example, in a sentence "I noticed a tall bush between two cedars."

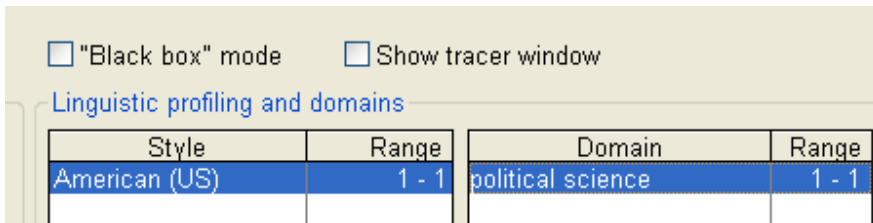


Now it's time for a different type of perennial plant. Let's copy an average newspaper headline, "**Bush made a speech about American policy in Iraq**". Carabao seems to understand that it has little to do with vegetation:

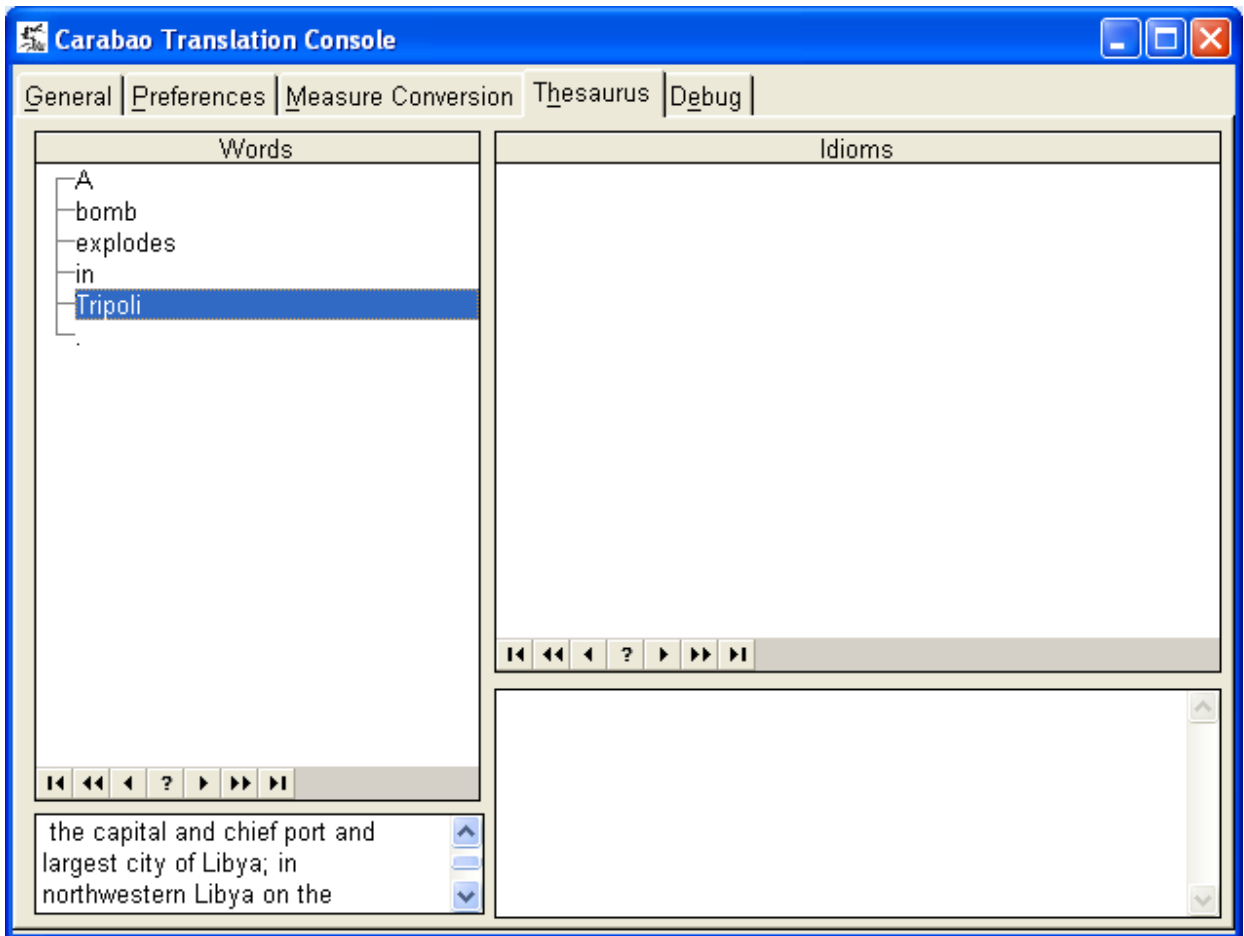


Note that there is no way of knowing in the current context whether it is George H.W. Bush or George W. Bush we are talking about (maybe it's the former?). It is possible, however, to enter more data manually and deduce the correct person in some cases.

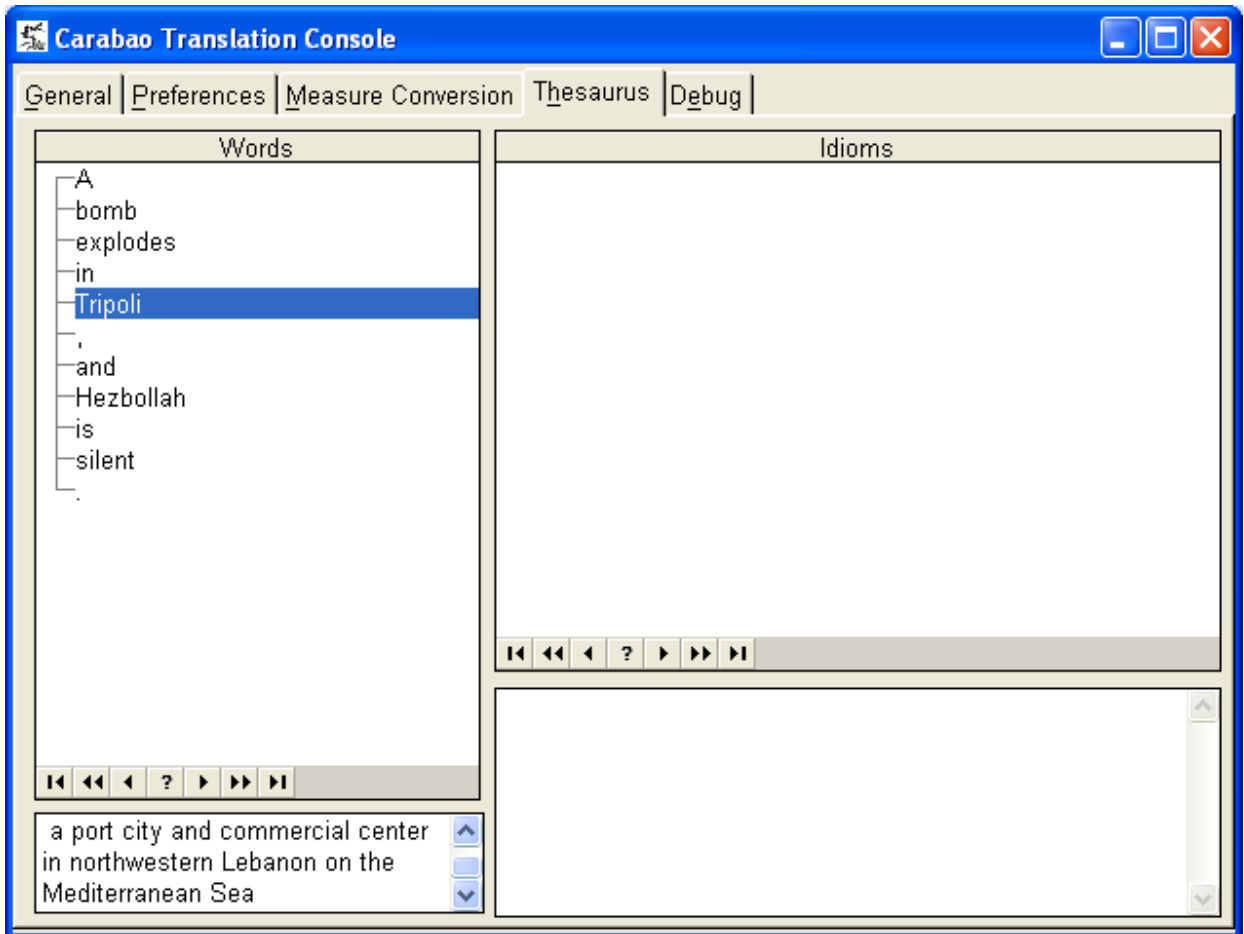
All seems correct in the table of detected domains:



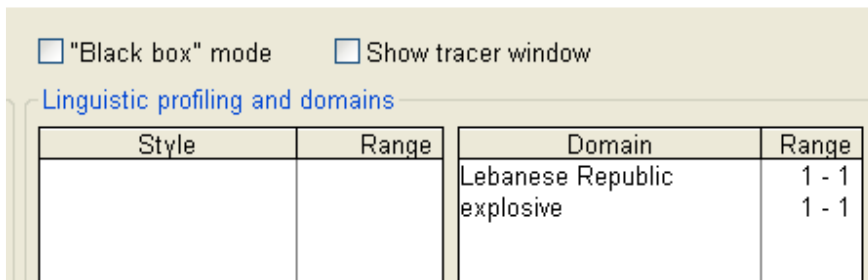
While we're talking politics, let us examine Carabao's reasoning capabilities. There are at least two cities in the Middle East called Tripoli. One is the capital of Libya; another is a city in Lebanon. The capital of Libya is a default when given no additional information:



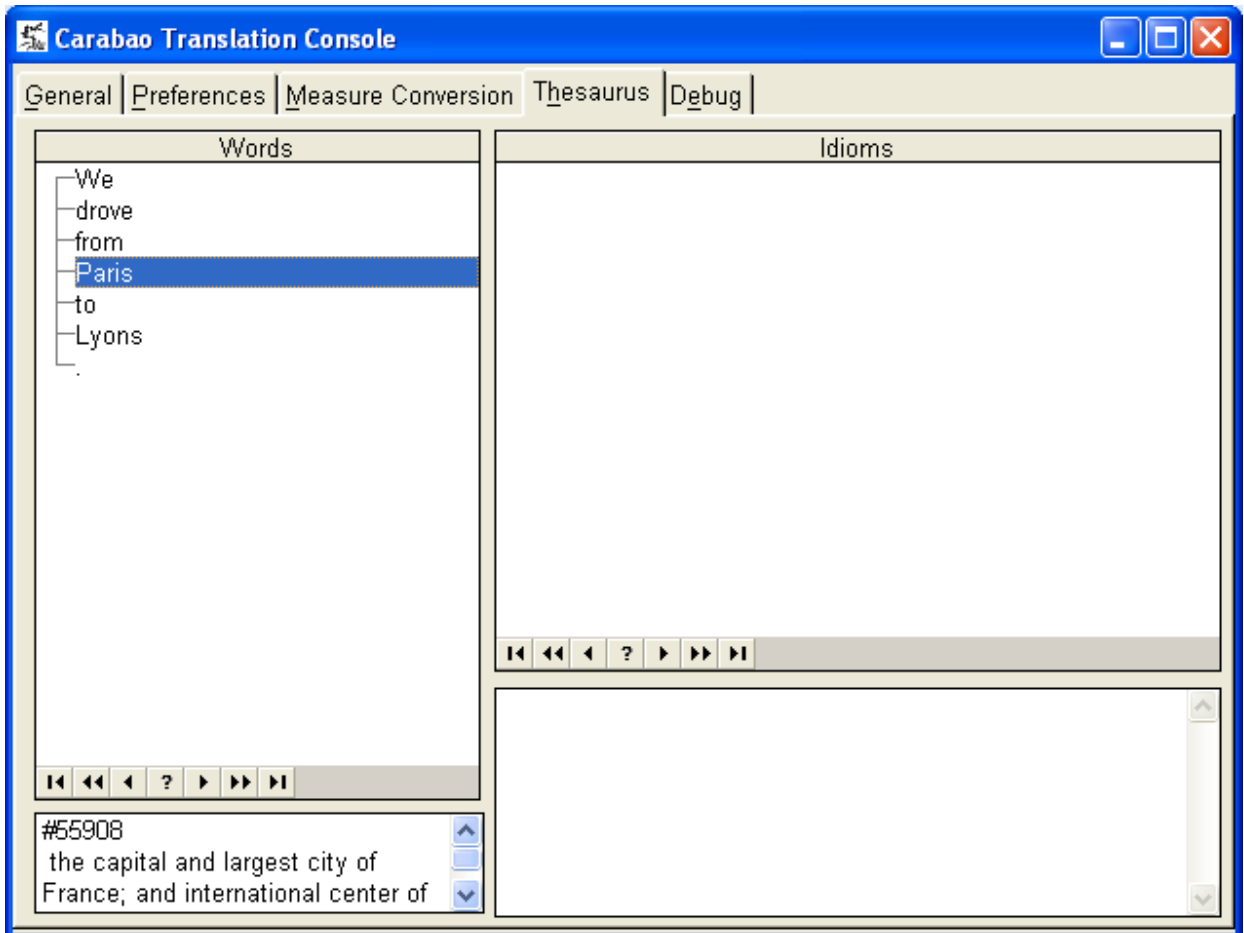
But what would a human deduce from a sentence like “**A bomb explodes in Tripoli, and Hezbollah is silent.**”? If he is not well-versed in Middle Eastern politics, then, probably, nothing. But if the human knows just a bit about Lebanon, s/he would agree with Carabao's conclusion:



And the list of detected domains demonstrates how Carabao arrived at this conclusion:



Geographic locations can be ambiguous. There is Paris in France, and there is Paris in Texas. If we're going from Paris to Lyon, then apparently we are in France:



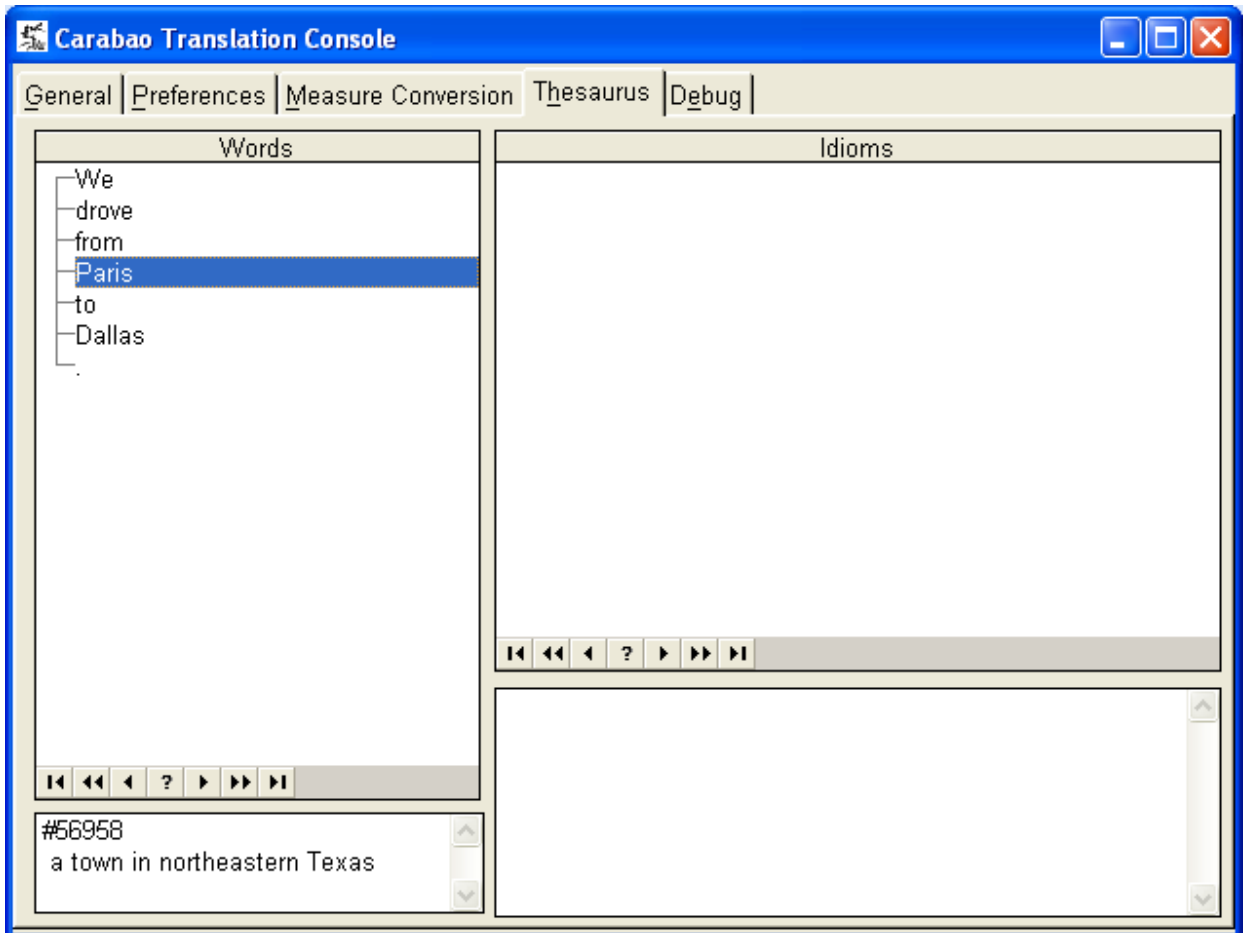
And the domains are:

Linguistic profiling and domains

Style	Range	Domain	Range
French	1 - 1	French Republic	1 - 1

'Style' is marked as French since Paris and Lyon are perceived to be words of French origin.

If we are going from Paris to Dallas, odds are that we are in Texas. Note that this can be wrong, but this is Carabao's best guess.

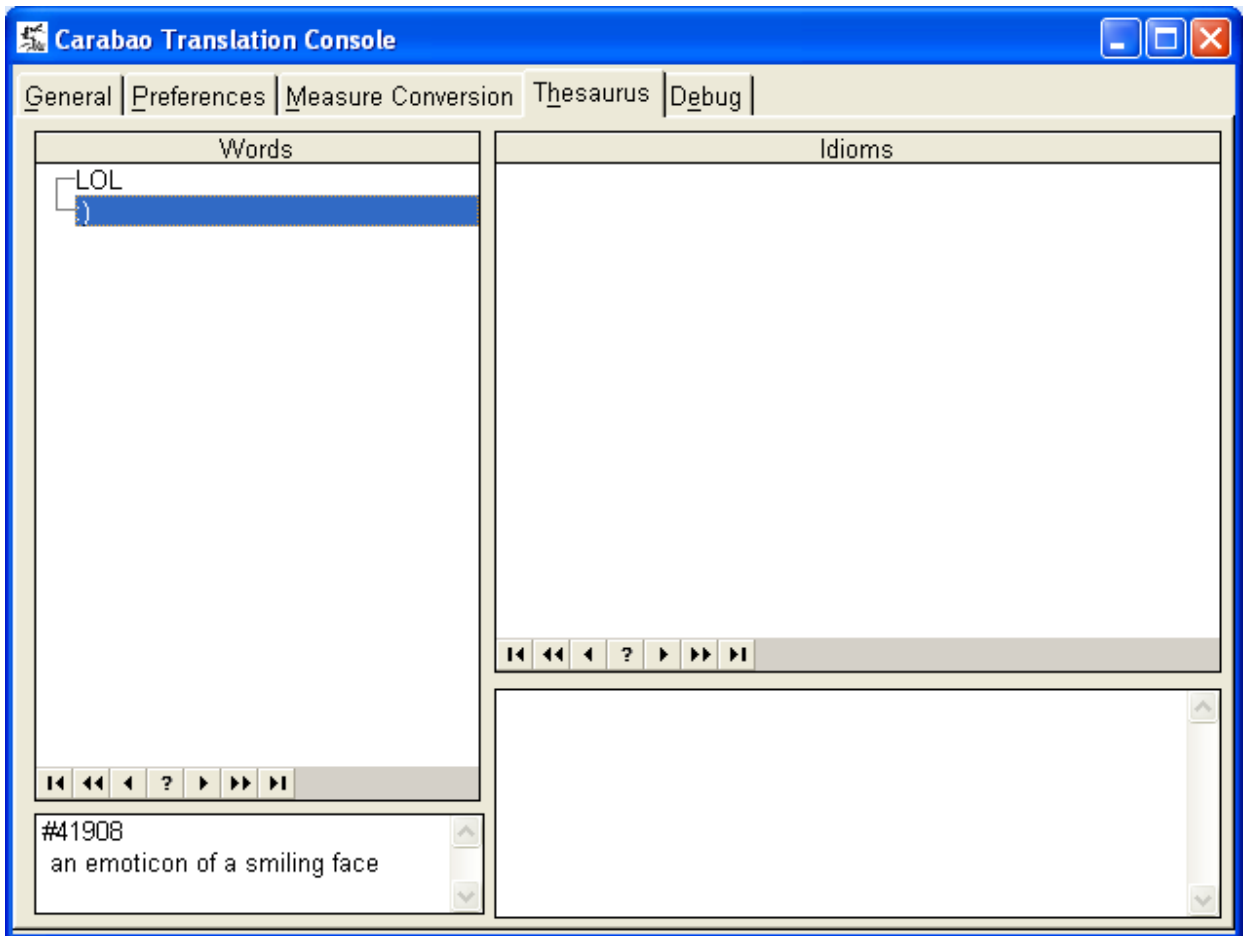


The domain list is 100% correct – 'Lone-Star State' is Texas:

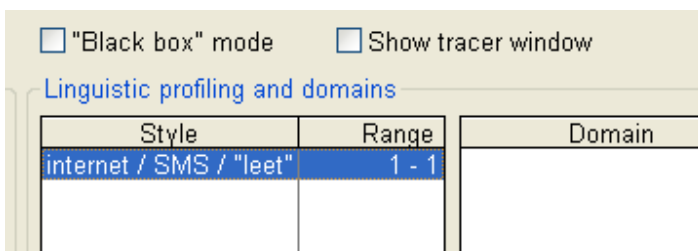
Linguistic profiling and domains

Style	Range	Domain	Range
		Lone-Star State	1 - 1

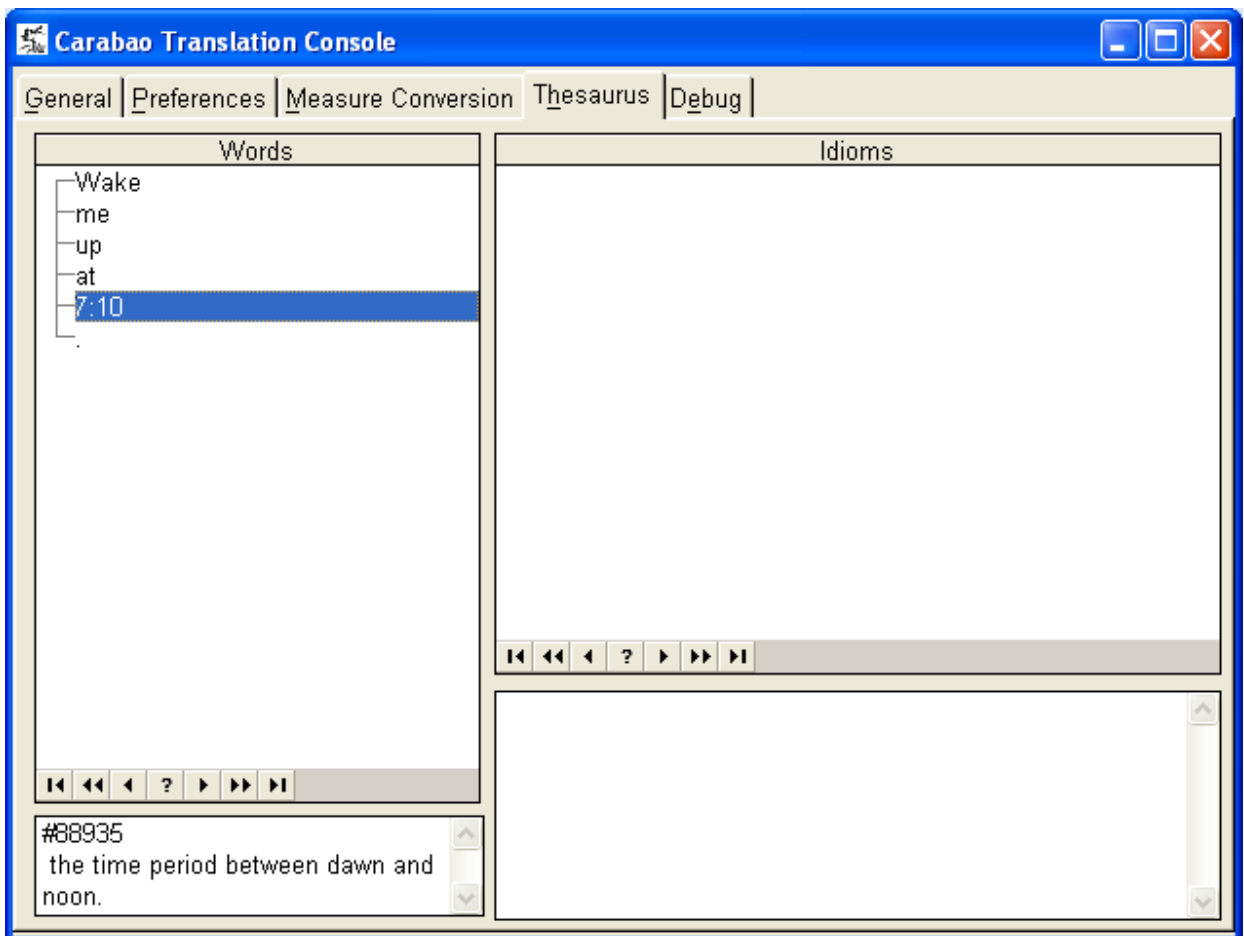
How about non-standard language, such as “internet leet”? Carabao’s data structure allows for the recording of such irregularities easily, so correct recognition is not surprising:



With Carabao’s powerful named entity recognition, it is able to recognize all varieties of smileys, as well as websites, emails, fractions, amounts of money and more. Linguistic profiling tells us that the author is, apparently, “kewl”:

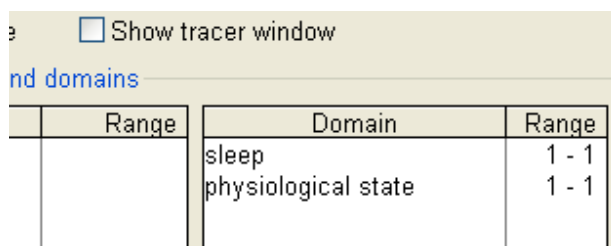


Associating patterns with semantic concepts lets us accomplish more fun stuff. For example, not only does Carabao identify clock time, but it also determines which part of the day is being referred to:

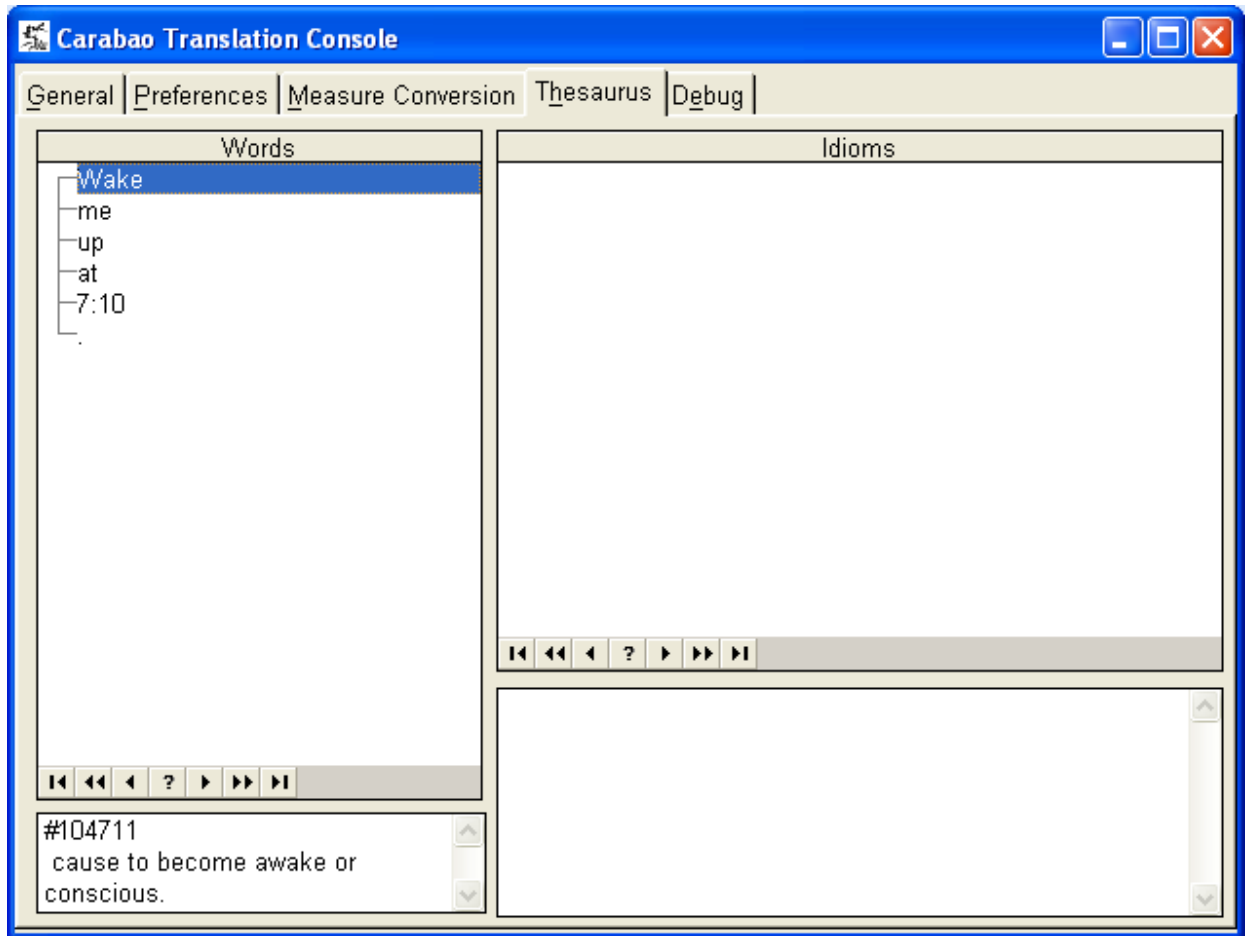


(No, we did not enter the every minute of the day. You can try adding seconds or milliseconds.)

And the fact that it's morning gives Carabao a hint about the domain:



and what the word “wake” means:



Summary

It took us a while to raise Carabao. And while we are proud of its achievements and can go on for hours, we have to admit: Carabao is not omnipotent. There are certain situations when even humans are not sure about the correct sense. Carabao is yet to reach the human level of understanding, so the usual disclaimer applies, even though it seems to be more robust than most other systems.

Perhaps another thing worth mentioning is that once a new language is added, the analysis will work for it too – because semantic links are language-independent. A truck is a vehicle, a fox is a mammal - in all the languages, and whether it is English, or German, or Russian, it does not matter.

Contact Details



Digital Sonata Pty Ltd
info@digitalsonata.com