LinguaSys
UNDERSTAND | ANALYZE | TRANSLATE

The Missing LNKR™

By Brian Garr, CEO

Almost all software companies complain about the same enduring problem: They designed their software to work on English input, but increasingly have to deal with other languages. After-all, only a fraction of Internet users speak English as their first language; growth is often faster in non-English markets; and the importance of international business means enterprises increasingly demand *multilingual* solutions.

Then there's a related challenge – surging demand for Natural Language User Interfaces (NLUI) and personal assistants, which means enabling computers to understand human language.

LinguaSys has developed some ground-breaking technology to solve these challenges. Our goal is nothing less than to provide software vendors around the world with a universal tool that will allow their software to understand any LinguaSys-supported language. Not even leading technology providers such as IBM, or Apple, have been able to solve these challenges on their own. They now realize that simply localizing their UIs (user interfaces) into other languages is usually not enough. The core software needs to be able to process other languages.

Imagine a world where your great application can understand client input in 17+ languages.

Remember the childhood game of telephone? As each kid whispered the message to the next child, it lost more and more meaning. This is what happens in current MT (machine translation) processes. That's because even the best MT – we know, because we're an MT provider as well -- has significant error rates. Even when the translation is exceptional, there is often an important loss of meaning.

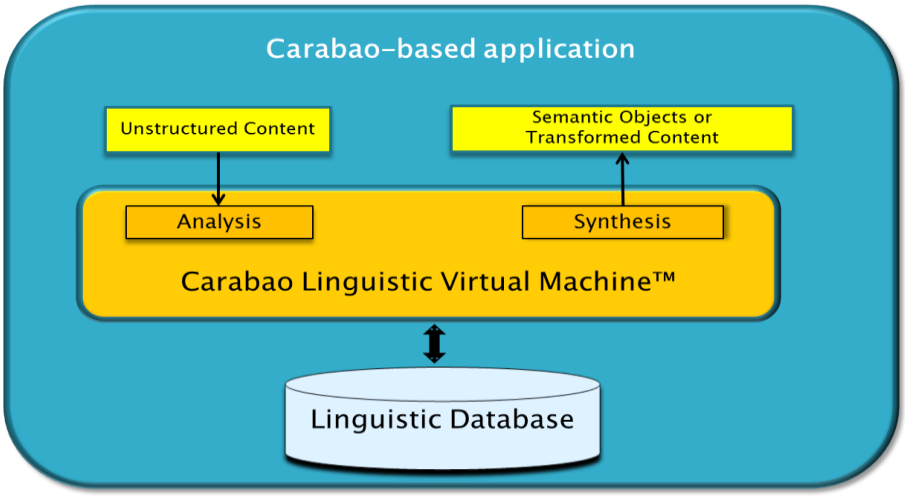
For example, German names of trade people may have gender, while English ones don't (*Bäckerin* is German for "baker" in feminine, that is, a female baker). Japanese and Korean have multiple degrees of honorifics and mood-reflecting particles with no adequate translation in English. There are numerous ambiguities in all languages (for example, is that "tank" you're discussing a military vehicle, home for fish or jail cell for drunks?)

Rarely is there a one-to-one relationship between a word and its translation. This is a major problem for MT solutions, which are generally based on statistics with a fundamental focus on English as part of the language pair. In otherwords, they review large amounts of text and then translate by weighing how often one word is associated with another.

Such algorithms that look for specific words in English often misfire when the content is not covered in their statistical map.

Our missing LNKR system does not focus on "words," but instead extracts and processes the "concepts" that are common to *all* languages. A global "concept map" – or "knowledge representation" -- is what LinguaSys has developed. It's a semantic-based (versus statistical) language-neutral (versus English-based) technology… It works with our patent-pending Carabao

natural language processing engine and LinguaSys proprietary language models to generate those missing links.



We can now create a *Language Neutral Knowledge Representation*, or LNKR, of the original, native content, which can be in English, Japanese, Malay or any of the 17 languages we now support. The system doesn't care. It creates a knowledge representation that converts unstructured content in different languages into a set of well-structured entities, uniform across all supported languages.

Here's how it works: the client application sends the input text (say a Tweet, or customer service request) to our core software – called the Carabao Linguistic Virtual Machine -- which consults the Linguistic Database to analyze and gather linguistic information on each word in the text – for example, morphological, contextual, and syntactic information. The system then creates the Language Neutral Knowledge Representation (**LNKR**), for consumption by the client application.

This is also extremely powerful for NLUI and personal-assistant solutions. **LINKR** allows NLUI scripts to be produced at a fraction of the time and cost of other solutions. Plus once it's built, it works in all 17 languages. There's nothing else like it.

Our *knowledge representation* crosses all languages. It is, in effect, a representation of the key concepts used by all human beings across all languages.

It avoids the "telephone-game" problem, because there is no machine translation to other languages in the processing.

At LinguaSys we're fans of the Bard. He wrote the famous lines: "What's in a name? That which we call a rose, by any other name would smell as sweet."

We have built a knowledge representation that understands the "concept" of "rose," and thousands more, in every language we support.

LNKR and the software behind it also avoid a critical problem in some other approaches. It does NOT use English as the frame of reference. All natural languages are ambiguous, and English is one of the most ambiguous. Even in cases like location names, it is not guaranteed to be well-aligned. For instance, take Georgia the country and Georgia the state (let alone the name Georgia); while in languages like Russian the country and state are two different words, in English it's the same word. That means a semantic network based on English will have to create different branches for these two concepts, and many more -- just in Russian -- defeating the purpose of the taxonomy.

Languages are ambiguous. LNKR is not.

Our system works at the word and compound word level to eliminate ambiguity, "understanding" a word by evaluating options that include: one to one match (which is rare); one to many; many to many; and one or many to none. In a many to none scenario, we assume the source word is a loan word.

We refine the words down to unique, basic concepts. In **LNKR**, if you say, T=12345, it identifies a specific sense of a concept **across all languages**. These indexes are unambiguous and create a backbone for semantic mark-up of text, and even for a larger semantic web.

Let's take a look at the following sentence:

My cellular phone is perfect.

This is a simple declaratory statement with low ambiguity. But even here there is some complexity not apparent at first blush. In the picture, we show versions of that sentence using the synonym "mobile" instead of "cellular" and in Spanish, Russian, and Simplified Chinese.

Language	Input Sentence	Language-neutral representation
English	My cellular phone is perfect	A=22$T=301%A=177232$T=**26300**$Y10=2$Y12=FOSE$Y13=LOW%A=113$T=308%A=6117$T=**5309**$Y13=MED%
English	My mobile is perfect	A=22$T=301%A=2197726$T=**26300**$Y1=JAR$Y13=MED%A=113$T=308%A=6117$T=**5309**$Y13=MED%
Spanish	Mi móvil es perfecto	A=2082735$T=301%A=2197727$T=**26300**$Y1=JAR$Y13=MED%A=2082213$T=308%A=2091410$T=**5309**$R5=M$R25=ACT$Y13=LOW%
Russian	Мой мобильник идеален	A=2476403$T=114268%A=2903406$T=**26300**$Y1=TALK%A=14152983$T=**5309**%
S. Chinese	我的手机很完美	A=3078604$T=301$Y9=L%A=3225370$T=**26300**$Y13=LOW%A=3222208$T=427%A=3217686$T=**5309**$Y9=L%

The graphic demonstrates that regardless of the input language, the **LNKR** produces a common output of all 5 utterances for processing.

The output looks like gorp when viewed at this level, but it contains all the data needed to be processed by multiple applications, from analytics to NLUI (natural language user interfaces) or even MT.

As an example, notice the "$T= statements" point to specific concepts in the language model. "#26300" is always the concept of a "cellphone", whose synonyms in English include cellular phone, mobile phone, cellular telephone, mobile, cell phone, cell, and cellular. If a popular new slang term for cellphone emerges tomorrow, we can just add to concept ID #26300.

So how does **LNKR** become the famed babelfish of "The Hitchhiker's Guide to the Galaxy?" Let's take a deeper look into what an outside application developer has available in the LinguaSys bag of linguistic goodies that will help their application consume the LinguaSys **LNKR**.

Keep in mind that it's not possible here to show *all* of the information available through the Carabao APIs. Our focus is to show the conceptual mapping of utterances to **LNKR** and how that leads to the knowledge for a software process to react.

Here is a sentence and the **LNKR** output:

I need to fill the tank of the tank.

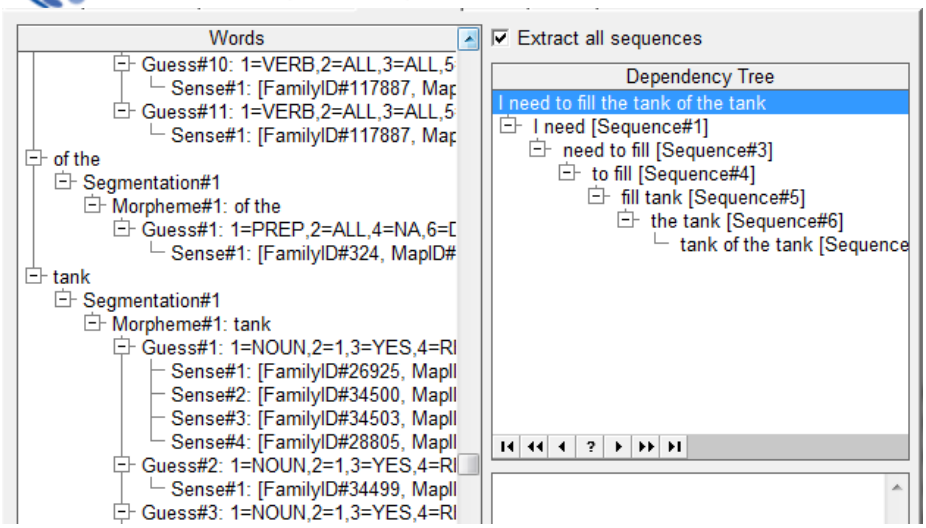
A=1$T=301%A=136357$T=112358$Y13=MED%A=452$T=323%A=128399$T=108811$Y13=HIGH%A=117$T=309%A=3509686$T=28805$Y1=JAR$Y13=LOW%A=2761459$T=324%A=37343$T=34500$Y13=MED%

If you peak inside our linguistic database, you can see that the concept behind T=28805 is the gas tank and #T=34500 is the military tank.

Language	Family ID	Lemma / Description	Note
English	659506	Tanjuubashi	location
English	26925	tank	
English	34499	tank	
English	34500	tank	a military tank
English	34503	tank	
English	83371	tank	
English	107475	tank	"tank animal refuse"
English	117887	tank	
English	28805	tank	gas tank
English	34503	tank car	tank car
English	34504	tank circuit	tank circuit

LinguaSys

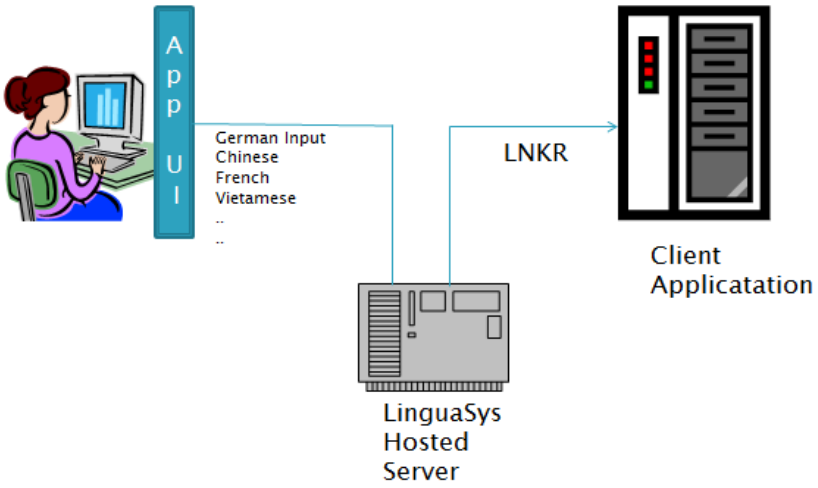
UNDERSTAND | ANALYZE | TRANSLATE



In addition to the lexicon of concepts and terms, our unique model for each separate language has a set of affixes associated with their inflection pattern, some shallow parsing rules, instructions how to handle unknown words, and syntactic structures. In the graphic above, you can see how the engine looks at every possible sense of the word (note different Family ID values for each entry) and then, through a system of voting and weighting, chooses the correct meaning word for this instance.

The ability to know so much about the meanings and concepts in an utterance, heretofore unavailable from statistical systems, will open up many exciting opportunities for software companies large and small to embrace a multilingual audience.

Think how this power might impact your business… Let's look at a potential deployment model. Once your application understands **LNKR**, it understands 17+ languages using the exact same back end logic. When we add an 18th language, there is NO CHANGE to your application.



We will shortly be launching **LINKR** as a hosted Web service. You can send your input and get back conceptual output your software can effectively process…across all 17 languages with more coming

LinguaSys, Inc.™
One South Ocean Blvd, Suite 306
Boca Raton, FL 33432

all the time. We can also easily add specialized vocabulary, such as medical terminology, to our models for those with such needs.

If you are interested in working with us on the Beta of our systems, please let me know right away. We are starting with three partners in the immediate future. Of course, we're quite happy to chat with you about our solutions for other language challenges, from multilingual data analytics to NLUI.

We live in a world where language is becoming more and more important. Remember the old song: "You like potato and I like potahto. You like tomato and I like tomahto. Potato, potahto, tomato, tomahto. Let's call the whole thing off."

We want to make sure your users, customers and clients never call off – or never even start -- relationships with you because of language issues.