Imagine you are standing in the middle of a busy airport, Heathrow or Dubai, everybody is talking. A lot of people, many conversations, many languages. People talk about their plans and experiences—where to stay if you go to Venice, which shoes are just simply awesome, what is the hottest play in town, why is this mobile phone the ‘must have’ gadget or why you better not buy this awful thing. Most of the time they are just chatting and by doing this they generate an enormous amount of information and feedback.

You listen to all those words but you don’t understand their meaning. Actually the Voice of the Customer (VOC), delivered via Social Media, E-mails, Transactions etc is like this airport. And the place is rapidly transforming from a big airport to a huge intergalactic hub. Between 2010 and 2020, the total amount of data will grow 50 times to 40,000,000,000,000 gigabytes. Nearly all of this data will be unstructured by 2015. This means that without text analytics nearly all of the VOC value would be lost.

The goal of text-analytic systems is to extract structured, actionable information from such conversations, from all markets around the world. Let’s create a mini case study.

A global mobile provider is fighting a dangerous increase of churn which raises specifically in Eastern Europe and Asia. They need to know why. And they need to know it now. Unfortunately, most of today’s systems focus on a handful of languages, mainly English. However, English speakers represent less than 30% of internet users worldwide, and this share is rapidly diminishing.

Moreover, as several reports show, existing systems still feel like rocket-science to their users. The API (Application Programing Interface) and terminology is overly complicated with a steep learning curve.

Our answer to these challenges is The Interpretor™, an open and flexible platform for analysis and interpretation of text, that can be characterized as:

- **Massively multilingual.** The system will support tens of languages, both large ones such as English and Spanish, and less commonly used ones such as Czech or Croatian.
- **Clean API.** The powerful API allows to integrate our system into customer’s workflow. It is simple, clean and unified across all supported languages.
Multilinguality

What sets our system apart from the rest is the unprecedented number of supported languages. And by supported we mean providing full-scale services, not just language detection as offered by some other systems.

Modern systems analyzing English are very accurate. However, they rely on resources such as large lexicons and manually categorized documents, creation of which required a substantial amount of expertise, time and money. For most languages and dialects, such resources are currently not available, for many they will never be.

To overcome this problem, we have developed a resource-light cross-lingual method. It uses a cascading combination of several statistical models, including models of related languages. We have been able to achieve accuracy comparable to the state of the art systems but with only a fraction of cost, labor and (language-specific) linguistic expertise. Because of that we can rapidly expand the number of analyzed languages. This being said, we have the confidence to say that our system is destined to be one of the best on the market.

Customer Benefits supported by Features

• When you submit a facebook or twitter post to our system, we first determine its Language and Topic (is it about a movie or a cell phone?) to route it to the right set of modules.
• Often the text needs to be “cleaned” because many posts are made on the go or in a hurry, containing typos and ad-hoc abbreviations.
• Then, the general Sentiment of the document is detected: Was the movie good? Is the cell phone worth buying?
• Our Entity Recognition module identifies the mentions of Brad Pitt, Samsung, or New York; correctly identifies that LA and Los Angeles are the same thing; tells you whether the author thinks highly of Steven Spielberg or Google and what their Wikipedia page or New York Stock exchange code is.
• And then come the Aspects. While the overall sentiment of the post is important, often a product or movie is not clearly bad or clearly great.

All of the above is highly customizable, the customer can specify sophisticated range of parameters, including the possibility to use their own terminology lexicons, labels, etc.

<table>
<thead>
<tr>
<th>Roadmap</th>
<th>Aug 15</th>
<th>Sep 15</th>
<th>Oct 30</th>
<th>Dec 31</th>
<th>Q1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theme</td>
<td>Version 1.0</td>
<td>Stabilization</td>
<td>Features</td>
<td>Languages</td>
<td>Speed/Robustness</td>
</tr>
<tr>
<td>Features</td>
<td>Lemmatization</td>
<td>Sentiment</td>
<td>Language detection</td>
<td>Topic detection</td>
<td>Text cleanup</td>
</tr>
<tr>
<td>Languages</td>
<td>Czech</td>
<td>English</td>
<td>Portuguese</td>
<td>Slovak</td>
<td>Polish</td>
</tr>
<tr>
<td>Interface</td>
<td>Webservice</td>
<td>Rapid miner</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input</td>
<td>Raw text</td>
<td>Twitter</td>
<td>Facebook</td>
<td>Generic web page</td>
<td></td>
</tr>
</tbody>
</table>

Geneea

We took the best from academia knowledge and the best from business approach, put it together and created Geneea and The Interpretor™.

If this seems rather pragmatic to you, you are right. It is what we believe in. It is the way we work at Geneea. It is what synergy means to us.

1 approx. ninety percent
2 See for example: Text Analytics 2014: User Perspectives a market study by Alta Plana