# Frequently Asked Questions

## Why language white papers?

The benefits offered by Language Technology (LT) differ from language to language and so do the actions that need to be taken. The META-NET white paper series “Languages in the European Information Society” reports (a) on the state of each European language and (b) on the state of each European language with respect to LT. It also explains the most urgent risks and chances. The series will cover all official European languages. Every volume of the series deals with one language.

## Who do the language white papers address?

While there have been a number of valuable and comprehensive scientific studies on certain aspects of languages and technology, there exists no generally understandable compendium that takes a stand by presenting the main findings and challenges for each language. The META-NET white paper series will fill this gap.

Rather than offering an exhaustive description of ongoing LT research and available technology, the white papers aim at raising awareness for LT support by depicting the importance of LT for every individual language as part of the European information society. **Addressing non-expert readers such as politicians and journalists,** the white papers should **clearly illustrate the main ideas in a generally understandable way**. In other words, the whitepapers are not scientific papers and, thus, should not be written in a scientific style. Rather, a style should be used that one could also find in a well-researched expose written for a renowned newspaper or magazine.

## How are the language white papers structured?

The white paper consists of three major parts.

**Chapter 1, the introductory part,** emphasizes the importance of LT and describes the chances and challenges within the European context. This part is fixed for all languages and should not be modified.

The second and third part (chapter 2 and 3) of each white paper address the specific situation of the particular language and the LT support for that language, respectively.

**Chapter 2 presents selected facts on the language**, its language community, particularities of the language as well as cultural, political and economic developments influencing the status of the language.

**Chapter 3 presents core application areas of LT** to the reader in order to demonstrate how LT is involved in and able to improve many well-known everyday applications. It also describes **the language-specific situation of LT** by taking a look at the research and industrial environment and by rating the availability of tools and resources for the particular language.

## What needs to be done?

The white paper template consists of language-independent (LI) and language-dependent (LD) parts. In order to adapt the template to your own language, you need to tailor the LD parts. These are described in the following for each chapter of the template.

Chapter 1 is LI and requires no adaptation. It will be optimised by a professional editor soon and will, then, be inserted in all white papers.

Chapter 2 is LD and **needs to be completely rewritten for your own language.**

Chapter 3 is a mix of LI and LD parts:

> The introductory part is LI and should be used in all versions; slight modifications are possible.
>
> The sections on the core application areas are mixed LI/LD sections: first they introduce the essential components (LI), followed by the **details concerning your own language** (LD) such as available tools and resources, companies in your country working in the respective area, etc. The current (beginning of March 2011) version of the template does not contain the LD parts for German yet. The description of one core application area should be no longer than two pages.
>
> The sections on research and industry should be LD and **describe the LT research and industrial environment in your country**.
>
> The table at the end of the chapter should provide an overview of the situation of LT for your language (see the next question). In this table, you will have to **assess the applications, tools and resources available for your language with respect to the text of the chapter**. Anything appearing in the table should ideally have been mentioned and highlighted somewhere in the text before. On the whole, the estimated ratings should underline the core message that the availability of tools and resources is not good enough yet in order to establish a truly multilingual Europe based on Language Technology.
>
> A conclusion will **pick up selected facts from the text and table and interpret them**. This will again be a mix of LD and LI parts.

## How to fill in the table on tools and resources?

The criteria used in the table such as *availability* and *quality* are explained in the current document template. As the table talks about whole areas such as syntactically annotated corpora, you have to fill in the columns in a very pragmatic way, e.g., *quality* will relate to the best resources while *availability* will refer to all resources. For the German report, we asked two external experts (who we list as co-authors, of course) to provide their values and estimates. We merged these with our own values. The agreement between us was generally very high. We recommend you apply this procedure for your whitepapers, too.

**The most important goal of this table is not to provide an exhaustive and scientific chart or overview of the field. The table is meant to support messages.** The messages can be explained in the section that follows the table. Among the messages that the example table in the German white paper contains are:

- “While some specific corpora of high quality exist, a very large syntactically annotated corpus is not available.”
- “For German, a very large corpus exists, but it is not easily/cheaply accessible.”
- “Many resources lack standardization, i.e., even if they exist, sustainability is not given (i.e., concerted programs and initiatives are needed to standardize data and interchange formats).”
- “Semantics is more difficult than syntax; text semantics is more difficult than word and sentence semantics.”
- “The more semantics a tool has to deal with, the more difficult it is to find the right data; more efforts for supporting deep processing are needed.”
- “Standards do exist for semantics in terms of world knowledge (RDF, OWL, etc.); they are – however – not easily applicable to NLP tasks.”
- “Speech processing is currently more mature than NLP for written text.”
- Etc.

As you see, the table serves as a summary of the LT-part of the document and it can be used to explain multiple key messages along several dimensions and comparisons.

Please note that even though the table seems to be, at first glance, highly complex, it is quite the opposite: it is a *simplified* representation of our field. The numbers in the current version for German are a first draft and need to be refined in a few cases. Based on our own experience with the table, experts in the field are indeed able to fill in the numbers from 0 to 6 – please try this yourself for your language and do not feel intimidated by the high level of abstraction. Also, please always think about the importance of all columns, especially the first one: if a certain technology or resource simply does not exist for your language, then please pencil in a 0 (zero). After all, this 0 be an important message for our politicians and journalists: a certain technology or resource simply does not exist for a certain language which is why we need additional support to bring our language and its LT up to international standards.

In the medium to long term the table will be further simplified to arrive at even simpler, even more abstract numbers that can be easily compared, especially across languages. Please note that no information will be distributed outside our network before all parties concerned are happy with them – due to its high level of abstraction this is especially valid for this table as it is an important instrument.