COMPETITIVENESS AND INNOVATION FRAMEWORK PROGRAMME

ICT Policy Support Programme (ICT PSP)

Multilingual Web

Objective 5.3: Multilingual Web content management: methods, tools and processes

CIP-ICT PSP-2009-3

Grant agreement for: PILOT TYPE B

Project Acronym:	**MORMED**
Project Full Title:	**Multilingual Organic Information Management in the Medical Domain**
Grant agreement no.:	**250534**

Evaluation of Machine Translation Technology for the MORMED Platform

Deliverable D 4.2

WORK PACKAGE: 4

LEADING PARTNER: LTC

SECURITY CLASSIFICATION: PU

DATE 2010-11-05

VERSION 0.12

Document History

Version	Date	Modification Reason	Modified By
01	2010-10-12	First draft	Laura Canedo (LTC)
02	2010-10-22	Contents and revision	Dr. Adriane Rinsche, Laura Canedo (LTC)
03	2010-11-01	Revision and amendments	Dr. Adriane Rinsche, Laura Canedo (LTC)
04	2010-11-05	Final draft, to be circulated among project partners	Dr. Adriane Rinsche, Philip McConnell, Laura Canedo, Nadia Portera-Zanotti (LTC)

Table of Contents

Executive Summary

This document presents the work performed during task 4.3 of the MORMED project, which corresponds to the evaluation of machine translation technology.

After a short overview of the MORMED project and the LTC Communicator II real time or near real time automated translation capability envisaged in the introductory chapter 1, an overview of machine translation (MT) technologies is given in chapter 2, and both the rule-based, empirical and hybrid approaches in machine translation are described, together with an explanation of different uses and purposes of machine translation, as well as generic selection criteria.

In chapter 3, we explain the evaluation methodology used. Analysis criteria and requirements for machine translation as part of the MORMED project are explained.

In chapter 4, we present evaluation results leading to the choice of Systran for the translation between English, German and Spanish, together with MorphoWord Pro for the English-Hungarian language pair. A diagram shows the linguistic workflow within the translation layer of the MORMED platform.

In a first phase, Systran, a commercial Rule-Based Machine Translation (RBMT) system with some hybrid functionality will be customised and used for translation, combined with subsequent human post-edition in order to achieve high quality translation output. We envisage using Systran as a training system to generate sufficient data volume that will allow us to switch to Statistical Machine Translation (SMT) later on. During this phase, bilingual aligned texts will be created in order to feed the Translation Memories (TM) and Statistical Machine Translation system used in the more sophisticated workflows in phases 2 and 3.

Phase 2 introduces a TM system prior to the RBMT system in order to improve the translation output. All human post-edited material will be automatically fed back to the TM. By this means we will obtain an increasing volume of training data for subsequent use in the SMT system.

Once we have a sufficient volume of data (bilingual corpora consisting of millions of words), we will test the SMT system Moses within the LTC Communicator II translation workflow, subject to positive evaluation results after populating and testing Moses with large corpora. We expect this to result in better translation output in the fairly narrow domain targeted in the MORMED project. The use of Moses would also reduce the cost of the final product.

At this moment, it is not yet clear whether Moses can finally be selected to replace Systran in our translation workflow system.

1. Introduction

MORMED proposes a multilingual community platform combining Web 2.0 social software applications with semantic interpretation of domain relevant content, enhanced with automatic translation capabilities, fine-tuned for a specific domain. It will be piloted upon the community interested in the Lupus disease or the Antiphospholipid Syndrome (Hughes Syndrome). Such rare diseases involve a significant number of people, either as patients suffering from this serious health problem or as experts, general practitioners (GPs) and researchers focusing on confronting and alleviating their hazardous aspects. Despite the fact that numerous people could contribute to the community with their knowledge about and experiences in these diseases, existing sources of information are dispersed and rather difficult to find. Thus, the need for a common repository focusing on these diseases becomes evident. Adding to the problem of the disparity of information resources, there are considerable language barriers that need to be overcome between researchers, consultants, GPs and patients speaking and writing in different national languages, so that anyone can access the information that addresses specific interests and needs. Consequently, there is a need for straightforward contribution and instantaneous exchange of information, regardless of the language in which it is expressed.

The proposed platform will result from integrating two main components. The first is a Web 2.0 semantically enhanced Content Management System, OrganiK[1], which will provide the presentation, business logic and storage layers. The second component is LTC Communicator tool, a web-based multilingual eCommunication tool, product of the Language Technology Centre Ltd, further extended with text summarisation functionalities.

LTC Communicator tool is currently under redevelopment in order to meet the challenges of the future. The redeveloped version will be called LTC Communicator II and will provide all translation and text summarisation services necessary for the realisation of the MORMED multilingual system by combining machine translation capabilities with a rich translation memory and optimal post-editing and human translation, available upon request. OrganiK's content management workflow will be enriched by LTC Communicator's automated translation functionalities and, along with the online community-specific support, form the MORMED frontend.

LTC Communicator is a web-based multilingual eCommunication tool. Machine translation and translation memory technology can be combined with human post-editing by translation experts to enable instant translations in many different environments. The system combines established language technologies, such as interfaces to machine translation and translation memory technology, with an automation server to provide a very flexible solution to multilingual content translation

[1] OrganiK Research Project at http://www.organik-project.eu - Seventh Framework Programme, Research for the Benefit of SMEs, Grant agreement no.: 2222225

requirements. External systems and applications can interface with LTC Communicator II to supply content for translation and receive the translated result.

Figure 1 shows the architecture of LTC Communicator II. Machine translation tools will be integrated to build up the automated translation backend of the MORMED platform.

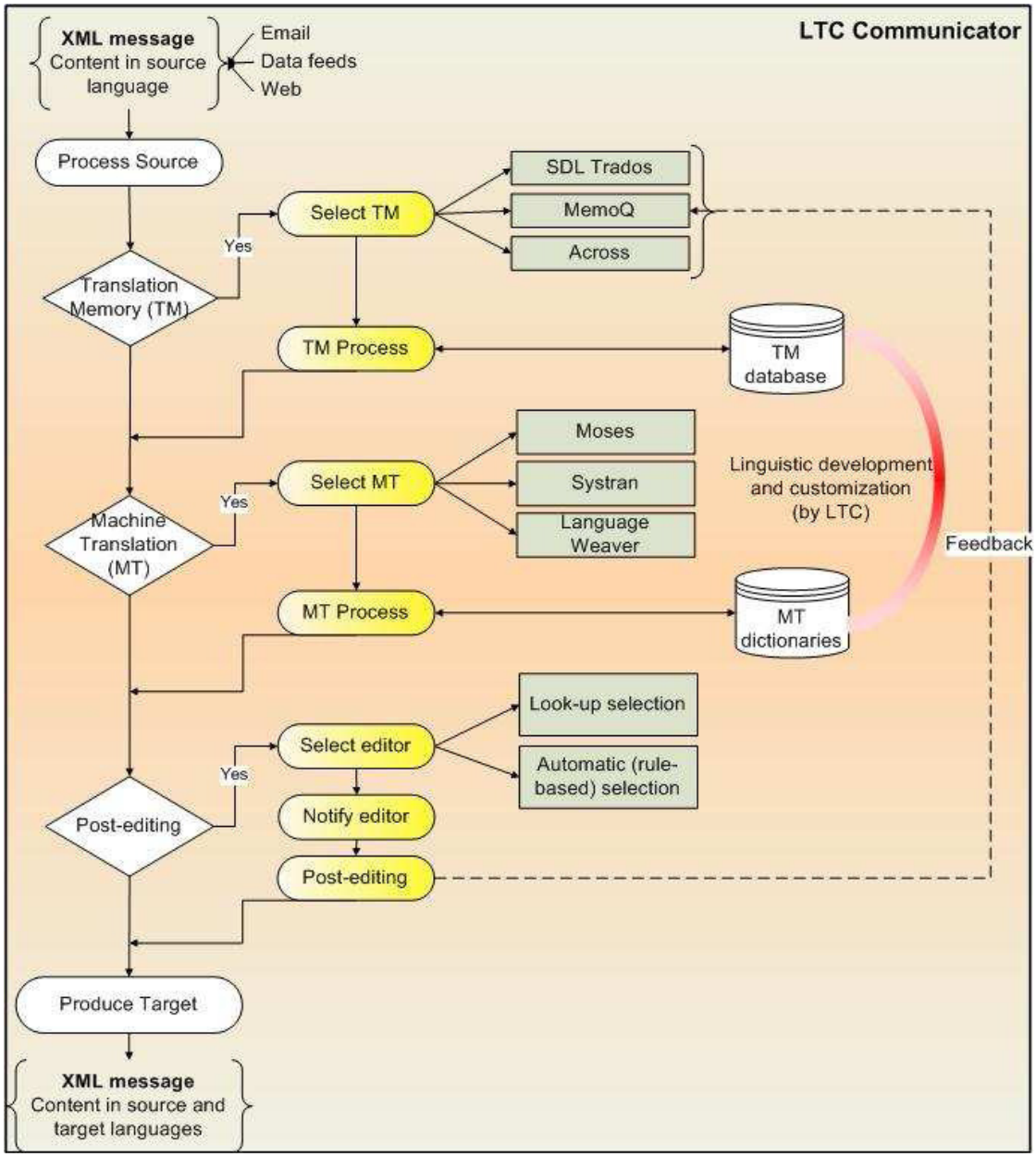

Figure 1 - LTC Communicator 2

Efficient machine translation trained to meet the needs of the medical domain and supported by interactive computer-aided translation and human post-editing will ensure that all content is seamlessly offered in any language and at high quality. Thus, new and innovative translation

methods, tools and processes will emerge, which significantly differentiate the service offerings of the MORMED service provider.

The multilingual component of the MORMED communication platform comprises four languages, ideally in all combinations: English, German, Spanish and Hungarian.

The automatic translation is carried out by means of a set of translation tools and an automation server routing translatable material through the different steps in the translation process and to human post-editing.

There are different types of translation software on the market, these will be described in detail in the following sections.

First of all, this document outlines the background of machine translation. Different approaches are described, together with an overview of current machine translation technologies.

The approach in the MORMED platform regarding the use of the machine translation technology will be flexible: ideally, a Translation Memory system will be combined with a Rule-Based Machine Translation system and/or a Statistical Machine Translation system.

As the selection of an appropriate machine translation system is an important element of the success of the final product, we will explain in this deliverable how a machine translation system has been selected, and the evaluation process adopted.

We will then explain the methodology and analysis applied: criteria and system requirements, constraints found, evaluation process, etc.

Finally, we will present the results together with a workflow of how a feasible translation process will emerge and how LTC Communicator II will be deployed.

2. Background

2.1. Machine Translation (Non Human Translation)

Machine translation, commonly known as MT, is an automated process by which computer software is used to translate text from one natural language to another.

The need for machine translation derives from several reasons. In the first place, we need to overcome language barriers. With globalisation, we need to increase the speed of multilingual communication at an affordable cost. The amount of new information to be translated increases exponentially with the collaboration of international teams, the exchange of personal information across borders and the development of new technologies and social networks, thus making it impossible for human translators to get it all translated 'in time', at affordable cost and acceptable quality.

There is also the problem of consistency. Humans tend to vary and use different words in professional contexts in order not to repeat themselves, but for the sake of consistency it is necessary to stick to previously agreed terminology. This can be achieved with multilingual technology. (Hutchins, 2005)

Machine translation is done by means of translation technologies. These comprise a range of computer applications which are designed to help or replace language translators. Products which help the translator are translator workbenches and terminology products; machine translation systems translate without human intervention. Owing to the imperfect nature of most machine translation output, humans often revise the output to remove errors. (Jane Mason, Adriane Rinsche, 1995)

There are different types of machine translation systems, which are based on different approaches, according to their design.

The **two main approaches** are rule-based and empirical. (Cameron Shaw Fordyce, Xavier Gros, 2007)

2.1.1. Rule-based approach

The **rule-based** approach (also called Rule-Based Machine Translation, RBMT) is based on linguistic knowledge of language, i.e., linguistic rules based on the syntax, morphology, semantics, etc. of a language. These rules are defined by human experts. This way, RBMT systems rely on linguistic rules and dictionaries.

The linguistic modules in machine translation systems are responsible for the *analysis* of the source text, *transfer* between two languages and the *generation* or synthesis of the target text. The analysis produces a complete parsing of a source-language sentence whereby all the words and lexical items in a sentence are reduced to their basic grammatical components. The output of the analysis stage is used to create the translation in the target language.

In the analysis and generation stages, most systems have clearly separated components for dealing with different levels of linguistic description: morphology, syntax and semantics. Analysis is therefore divided into morphologic analysis (identification of word structure: stems, endings, ...), syntactic

analysis (identification of sentence structure and parsing) and semantic analysis (resolution of lexical and structural ambiguities). These phases may also be applied to the generation stage.

RBMT systems are the most sophisticated translation technology linguistically, since they are designed to carry out the translation task without any human intervention. The system is solely responsible for the translation of the source language into the desired target language and therefore requires special programs, comprehensive dictionaries and collections of linguistic rules. This information needs to be complete in order for the system to generate the target version.

The term 'batch' is used to refer to the translation mode of machine translation products since there is no human intervention during the conversion from one language to another. Human intervention only takes place, if at all, after translation: errors in the machine translation output are manually corrected. This stage may not be required if only information is sought but it is essential if the translation process is carried out for dissemination purposes.

Since these systems provide full linguistic capability which maps source information to target information, they are developed strictly to translate between language pairs.

There are three approaches in RBMT systems: direct translation, interlingua and transfer. (Jane Mason, Adriane Rinsche, 1995)

2.1.1.i Direct translation

The approach used in nearly all machine translation systems until the late 1960s was the **direct translation** approach, as shown in Figure 2.

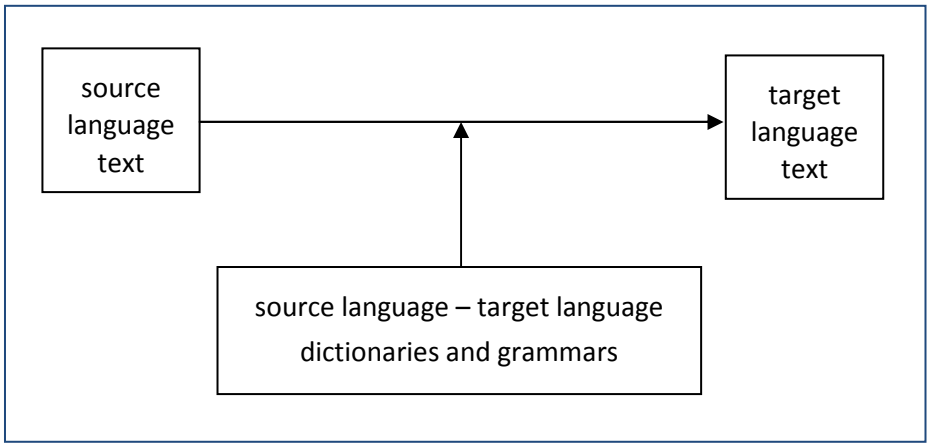

Figure 2- Direct machine translation

These systems were designed from the start to translate from one specific source language into one specific target language. They were 'word-for-word' translation systems, perhaps with some local word-order adjustment. Direct systems were limited to the minimum work necessary to do the translation for this single language pair. In these older systems, rules for analysis, transfer and generation were not always clearly separated. Some also mixed linguistic data (dictionaries and grammars) and computer-processing rules and routines.

An example of this approach would be the GAT system (Georgetown Automatic Translation), developed at Georgetown University in the US, which was the largest project in the early years of machine translation research. This project began in the 1950s with the purpose of translating Russian physics papers into English and is regarded as a typical 'first-generation' machine translation system. The Systran machine translation system has its origins in derivations from the GAT system.

Direct systems were linguistically very weak, offering only word-for-word translation resulting in poor quality output. Direct systems are no longer used commercially; they have been superseded by more sophisticated approaches.

By the mid 1960s, it had been recognised that machine translation could not advance much further without more sophisticated linguistic analysis. At this point, research into indirect translation systems began: the 'interlingua' and 'transfer' approaches. Systems of this nature are often referred to as 'second-generation' systems.

2.1.1.ii Interlingua

The **interlingua** approach to machine translation is shown in Figure 3. In these systems, source-text analysis and target-text generation are kept separate; conversion from one language to another is achieved via abstract 'interlingua', representations of meaning which are common to several languages. Translation therefore has two stages: analysing the source language to an interlingual representation and then generating the target language. Analysis and generation programs are independent; in multilingual systems, any analysis program can be linked to any generation program.

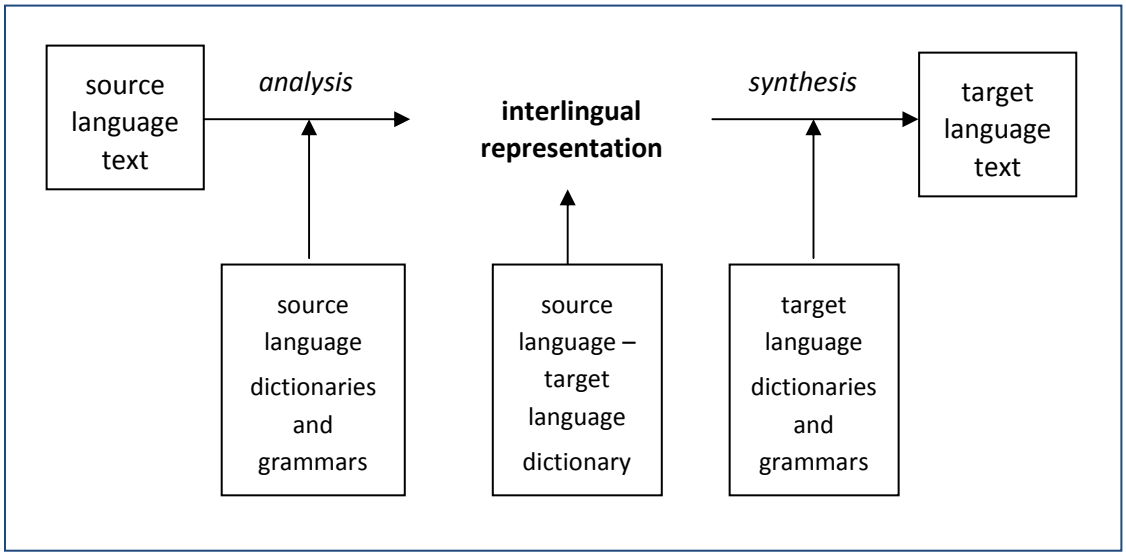

Figure 3 - Interlingua machine translation

Interlingual systems are considered theoretically superior to other approaches, since they require fewer language-dependent modules. However, the reduction in modules is generally negated by the effort required and difficulty involved in defining an interlingual representation for all languages.

An example of interlingual system is CETA (Centre d'Etudes pour la Traduction Automatique). This system was developed between 1960 and 1970 at Grenoble University in France, to translate mainly from Russian into French, and is still largely a research-based system.

2.1.1.iii Transfer

The third approach to machine translation is **transfer**, which is favoured in most current systems. The transfer module maps one language-specific representation of meaning onto another, as shown in Figure 4

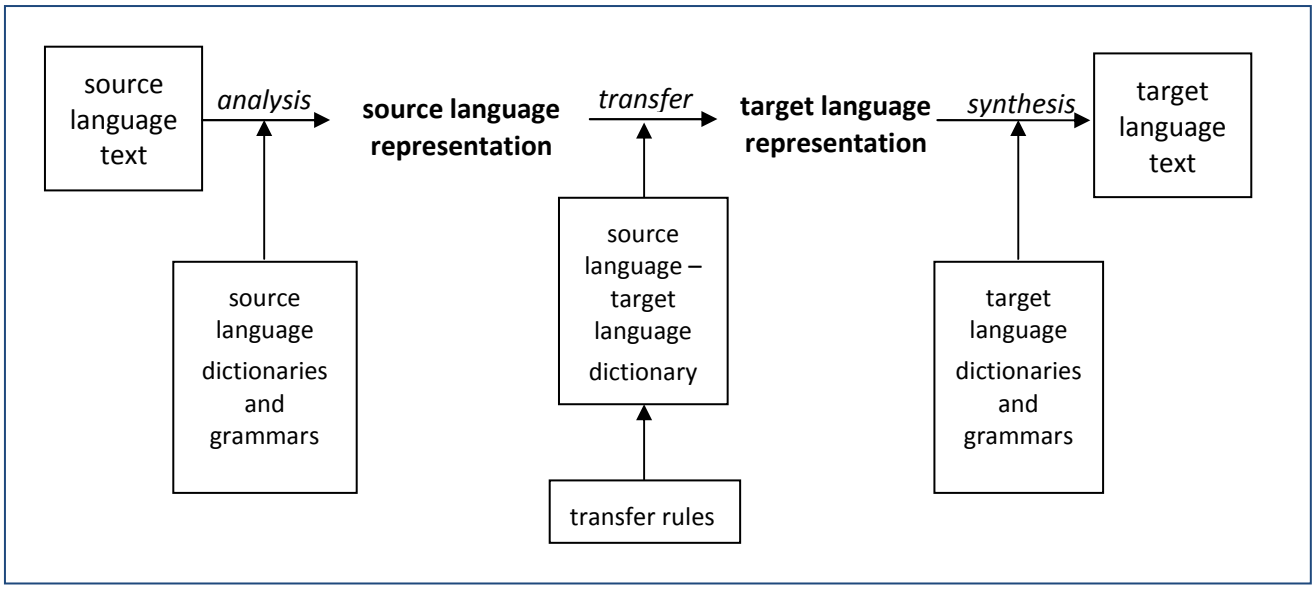

Figure 4 - Transfer machine translation

Rather than two stages via a single, highly abstract interlingual representation, there are three stages involving underlying syntactic representations for both the source and target language texts. The first stage converts the source language into 'deep' intermediate representations, in which ambiguities have been resolved without reference to the target language. In the second stage, these are converted into equivalent 'deep' representations in the target language. The final target language text is then generated. In transfer systems, analysis and generation programs are independent and are specific for particular languages. There may also be separate components to handle lexical transfer (selection of vocabulary equivalents) and structural transfer (transformation of source-text structures to equivalent structures in the target text). Differences between languages are dealt with by the intermediary transfer program. There are many operational examples of transfer-based machine translation systems, including Systran, Tstream (based on former Metal system) and Logos.

Until recently, the transfer approach is the most commonly used in current commercial systems, as it can be easier to implement than an interlingual system and can be derived from direct systems by augmenting the dictionaries and grammars. The interlingual approach requires complete resolution of all ambiguities and anomalies in the source text; in the transfer method, only those ambiguities inherent in the particular language need to be resolved.

Both interlingual and transfer-based systems are more modular than the older direct approach, so system components can be adapted and changed independently of each other.

2.1.2. Empirical approach

In the more recent and currently most 'fashionable' **empirical** approach, bilingual texts produced by human translators are used as raw data by the system for future translations. No linguistic language modelling is done.

Two types of systems pertain to the empirical approach: example-based machine translation and statistical machine translation.

2.1.2.i Example-based machine translation (EBMT)

Previous translations are stored in a database and offered as suggestions when new data needs to be translated. The idea behind EBMT is to take advantage of the similarity of a given input text to the example texts stored.

The texts stored are bilingual and aligned, i.e., a source text sentence (or segment, as they are called) is in one column, with the corresponding target sentence or segment next to it in a second column. These bilingual aligned texts are called **Translation Memories (TM)**.

EBMT systems are more thought of as an aid for human translators than as a machine translation system in itself. They are also called **CAT (Computer Aided Translation) tools**. Based on the concept of Translation Memory and the automated re-use of previously translated texts and sentences, the system offers the human translator a possible translation for their input text/sentence according to the translations stored in its database, and then the human translator has to choose and confirm the translation. Therefore, the output is not 'ready'; it is only a suggestion for the human translator, who finally decides whether or not at all to use a translation or parts of it. (Cocci, 2007)

Originally, CAT tools contained no linguistic rules and no ready-made bilingual dictionaries. However, the trend is to interface these systems with MT in order to improve and speed up the translation process.

The terminology databases are created and maintained by the translator with each translation and revision. In addition, term extraction tools can help automate populating term bases.

CAT tools are offered as translator workbenches (also referred to as translator workstations). They are integrated suites of tools designed to help the human translator, who is responsible for the translation of the text from one language to another.

The term 'interactive' is used to refer to the translation mode of translator workbenches, since the human translator creates the translation by interacting with the tools and on-screen information provided.

A CAT workbench offers a variety of tools in a single environment. Typical tools provided by workbenches include:

- multilingual translation memory databases, or techniques for file-based update,
- multilingual terminology databases or dictionaries (also called **term bases**),
- tools for building and managing terminology and translation memories,

- tools which look up terminology and incorporate the target terms into the translation,
- product-specific or standard multilingual word processor,
- tools for importing and exporting material,
- additional utilities for tasks such as managing files, counting words, printing and checking contents.

The above CAT tools also include a batch translation function to pre-translate highly repetitive text and keep manual manipulation to a minimum.

2.1.2.ii Statistical machine translation (SMT)

Translation examples (or translation memories, TMs) can be used to train a statistical translation model. This is just a mathematical model which performs a translation based on a statistical modelling of language.

There are *word-based* and *phrase-based* models. Word-based models (which are the most widespread) consider sentences as a combination of single words, ignoring the structural relations between them. Phrase-based models consider sentences as a combination of phrases or 'chunks', but not necessarily linguistic phrases. In both cases, the combination of elements is modelled purely statistically.

For the good performance of a statistical machine translation system, a large amount of data (in the form of aligned bilingual corpora) is necessary. These systems can therefore be customised by training them with corpora (bilingual and aligned) of the desired domain. It must be noted that acceptable translation quality requires an amount of training material of several millions of words per language within a narrowly defined domain.

All these types of products have several aspects in common. They need to recognise terminology in the source language and correctly identify equivalents in the target language. They also need to recognise sentences which have been translated before.

Additionally, they need to identify words and phrases in the source and target languages in order to produce accurately translated output or help translators efficiently.

The computational and human costs of the rule-based and empirical systems differ.

In the rule-based approach, prior modelling of the languages concerned is necessary, and this is to be done by a human expert. If wrong translations arise, new rules need to be created or inferred from the source and/or target texts in order to improve the translations. Therefore, both the computational and human costs are high.

In the statistical approach, systems for new language pairs and domains can be developed relatively easily, as long as enough data are available. An algorithm needs to be developed so that the machine can 'learn' starting from a set of new texts (bilingual and aligned). However, once the algorithm is created it is enough to feed the system with more and more aligned bilingual materials.

2.1.3. Hybrid systems

Recently, combinations of the two approaches are being developed in order to improve the translation output, in the so called **hybrid systems**. These systems combine advantages from both the statistics based and the rule-based systems.

The development of *rule-based machine translation systems* is time consuming and very expensive. The *statistical machine translation model* has high computational costs, but is robust in processing all kinds of texts. The linguistic quality depends on how much a new text corresponds to a narrowly defined domain. *Example-based machine translation systems* rely on bilingual texts which must be of high quality in order to result in good translation output. Having collected a good set of relevant translation memories, the translation output is of high quality, if a new text is similar to the information stored in terms of subject area and content. These systems become more useful in concrete domains like technical and medical translations, where the vocabulary and construction of sentences is more restricted.

Therefore, there is a trend towards the creation of hybrid systems which combine different approaches, making the most out of each individual system.

2.2. Configurations

(Jane Mason, Adriane Rinsche, 1995)

The first generation of machine translation systems were developed in the late 1950s and 1960s, when only mainframe computers were available and there were no high-level programming languages. Most programs for machine translation systems developed during this period were written in assembly code.

Today, machine translation systems are available in multi-user, client-server and web based environments or as single, PC-based products where the number of systems which support the latter configuration is increasing. Products also include different configurations for individual, personal users or for enterprises. Solutions for companies may vary depending on the required functionality: translation of texts and/or files, format choices, specialised resources (dictionaries and term bases) on certain domains, customisation capabilities (both in quantity and quality), etc.

In client-server and web based solutions, the principle is that all the linguistic data and programs are protected from general users and reside on a server, with the client having controlled access to dictionaries and other linguistic resources, if any. These are generally large-scale, centralised systems which run on powerful processors with sophisticated parsing, linguistic and lexical capability. They are aimed at multi-user environments and are relatively expensive.

The number of more affordable machine translation systems intended for single translators is increasing, as well as solutions for end users, who are no language experts.

Regarding translator workbenches, there is now a trend away from single user towards multi-user environments, satisfying the need to co-operate on a decentralised, networked basis on larger scale project, leveraging content available in a TM and dynamically generated by team members. Each translator is provided with a dedicated, powerful processor and an integrated set of tools for carrying out a set of tasks. In client-server and web based configurations, any terminology and translation memory databases are usually located on the server and are generally write-protected during translation to prevent the master databases from being populated with prior to final validation data.

Whatever the product configuration, compatibility with networks –both local area and internet– is important.

2.3. Usefulness and purposes of machine translation

Nowadays, it is possible to obtain an acceptable and understandable 'final' translation by a combination of translation tools and fine tuning.

Translation is a means of facilitating cross-language communication. Usability (mutual understanding) is more important than perfect results in some contexts.

Whether MT should be used is a decision which should be based not on whether the system produces 'real' translations, nor on whether it produces 'good' translations, but on the purpose for which it is required.

MT can be used for many different purposes, such as:

· *Gisting*. Sometimes it is enough for the users to just extract the essential content of a document. This unedited output is likely to be of a lower level of quality unless controlled language and considerable fine tuning have been done in advance.

MT systems work very well for this kind of use. Although poor, it is better than no translation at all. This use has grown rapidly and substantially with the coming of cheaper computer-based systems on the market and with the advent of the Google Translate function.

· *MT as a component of information access systems*. MT systems are more and more being integrated into different types of information systems (retrieval, extraction, management, etc.), as well as databases and other systems for data storage (data warehouses, etc.). This field is the focus of a number of projects in Europe that have the aim of widening access for all members of the EU to sources of data and information whatever the source language.

· *Localisation*. The internationalisation of enterprises is a common phenomenon nowadays. MT technology is often used for the translation of documentation and help files, localisation of user interfaces, maintenance of terminology, document management and version control, etc. (N. Puntikov, 1999)

· *Communication*. The demand for translations of electronic texts on the Internet, such as web sites, electronic mail and even electronic 'chat' lists is developing rapidly. In this context, the possibility of human translation is out of the question. The need is for immediate translation in order to convey the basic content of messages, however poor the input. The development of systems for spoken language translation is currently the focus of much research.

· *Professional purposes*. As explained previously, MT can contribute to the productivity and efficiency of the work of professional translators: special tools for terminology mining and an extensive set of individual user settings integrated in machine translation systems prove to be effective enablers for the professional translators.

· *Dissemination*. MT is used to reach a wider audience, overcoming language barriers. There is demand for translations of high, publishable quality. In this case, a combination of translation tools generally produce output which must be revised by human translators. Alternatively, the input text may be 'controlled' in vocabulary and sentence structure so that the MT system produces few errors that have to be corrected.

· *Quick update*. For many companies the need to translate updated documents quickly has been a pre-requisite to selecting a translation technology product.

2.4. Products available and selection of software

There are many different products on the market.

Every year, the European Association for Machine Translation (EAMT) publishes a *Compendium of Translation Software*, which can be used as a reference for an overview of the translation systems and the languages covered. (Hutchins, Compendium of Translation Software, 2009)

Companies and individual users can choose from a growing number of translation technology products, so selecting the most appropriate product is a complex procedure. As a general rule, issues which should be analysed include the need for translation technology, typical user environments, the materials to translate and their purpose, product suitability, user skills and the implementation process.

The time required to post-edit the output should be taken into consideration, if it is to be published. Therefore, generally speaking machine translation technology is really only suitable when:

- the subject domain and text type are restricted,
- the product is fully developed and customised,
- the desired quality level of the translated material allows for it,
- users require flexibility,
- it may be necessary to have control over linguistic aspects such as domain terminology,
- documentation is frequently updated,
- projects are decentralised,
- volumes of source materials are unpredictable.

3. Methodology and Analysis

In this section the procedure for the evaluation and selection of MT tools will be described.

The steps followed in the process are:

1. Definition of criteria, requirements and options.
2. Search for and selection of tools available according to the criteria expressed in 1.
3. Evaluation, assessment of tools' possible features not considered in 1, and study of the tool chosen, in order to explore its functionality in detail and how resources must be prepared to feed the system.

3.1. Criteria, requirements and options

First of all, it needs to be decided which are the criteria, requirements and options with regards to the translation functionality of the MORMED platform. They are listed according to their importance:

i. Language pairs

As per MORMED's Description of Work, the initial platform languages will be English, German, Spanish and Hungarian. The translation flow should run in all directions between all four languages. However, we are aware that this is a difficult option especially for Hungarian, for which there is a lack of MT tools. Therefore, it is likely that the translation flow will cover EN↔DE, EN↔ES and DE↔ES, on one hand, and EN↔HU, on the other.

ii. Maturity

The MT system should not be a research tool or prototype. It should be stable and its efficiency proved. Also, it should be easily available on the market and have a technical support service in case technical problems arise during the implementation and use phases.

iii. Price

In principle, we aim to use open source technology wherever possible in order to keep the cost of the finished system as low as possible. It will be considered as long as its maturity and efficiency have been proved. In case a commercial system is chosen, the price needs to fit the project budget.

iv. Customisation / Control

The resources (dictionaries, translation memories, term bases, etc) to be used by the MT system should be customisable, or at least the language technology experts in charge of implementing and fine tuning the solution should be able to have control over them.

v. Translation quality level

We are not looking for perfect results of the machine translation system. The aim is to give all users access to the information, regardless of their languages. In case a better translation is needed (for medical research purposes, for example), a human post-edition can be requested.

The quality level sought is understandability. During the course of the project, all translated text will be post-edited to control and improve output quality. We will continue adding language resources (dictionary updates and parallel text) to the system throughout the lifetime of the project.

vi. Purpose of the translation system

As stated in the Description of Work, the purpose of the translation functionality in the MORMED platform is to make all content available to all participants, to gather experience and make medical information accessible internationally and instantaneously to all platform users in the languages supported.

vii. Resources in the medical domain

The MT system should have in-built resources (dictionaries and bilingual aligned texts) in the medical domain in order to obtain quality translations. Rare diseases are a highly specialised field, so the MT system should be 'specialised' in this field as well. In case no predefined dictionaries or other resources are available, it should allow the user to customise the dictionaries and other resources in order to achieve high quality in the translation.

viii. Preparation of resources for MORMED

The customised resources to be fed into the MORMED system, such as dictionaries, term bases, bilingual texts, etc. need to be pre-processed prior to their introduction into the system. This preparation should be as easy as possible, without the need of deep knowledge of technical tools which would complicate the work both in terms of time and human effort, therefore resulting in an increase in costs.

ix. Input and output of the MT system

The input to the system will be either medical or general language with a medical flavour. In some cases (e.g. doctors or researchers), the language will contain professional terminology, whereas in other cases (e.g. patients) it might be less specialised, less technical. We also need to distinguish between informal content such as messages and questions on the MORMED platform as opposed to documents submitted for translation. Informal content will be translated on the fly without any post-edition, whereas documents will be subject to human post-edition.

In any case, it is not foreseen that there will be much input of colloquial language. The MT system should be able to cope with both more and less domain specific language.

The output should be understandable for the target user.

x. Post-edition of MT output

Ideally, the translation output should require as little human post-edition as possible. In order to obtain the highest quality possible, the systems will be continuously customised and tested for quality in an iterative approach. During the lifetime of the MORMED project all output will therefore be post-edited.

xi. Statistical Machine Translation

In principle, we are working towards using statistics based machine translation systems. These are proving to be efficient in restricted domains with large data volumes and in currently otherwise unavailable language combinations. Moses, a statistical machine translation system, meets these two conditions and is therefore taken into consideration.

xii. Integratability

The MT system should be easy to integrate into the MORMED platform.

3.2. Search and selection of tools available according to the criteria expressed in 3.1

The most recent edition (January 2009) of the *Compendium of Translation Software* published by the European Association for Machine Translation (EAMT) (Hutchins, Compendium of Translation Software, 2009) has been used to extract information regarding the different translation systems available on the market.

In terms of language combinations and machine translation quality, the three machine translation systems considered from the beginning were Moses (SMT), Systran (RBMT) and MorphoLogic (RBMT). The first two for the English-Spanish-German and the latter for the Hungarian-English-Hungarian language combinations.

Moses was taken into consideration due to the high quality that the emerging SMT systems seem to show when they are trained with a large amount of material and the domain is narrowly defined. The topics of the MORMED project are the Lupus disease and the Antiphospholipid Syndrome. These are rare diseases (i.e., narrow domains) for which the information available in parallel for multiple languages is initially scarce. Therefore, a SMT system will by definition result in poor translation output.

On the other hand, Moses is open source technology, which can keep the costs of the project and the cost of the platform after the end of the project at a low level.

However, due to the poor translation output when trained with small amount of bilingual texts, Moses had to be discarded as the MT technology used within MORMED for at least the first phase of the project.

Other free high quality translation services like for example Google Translate had to be discarded as well because of confidentiality requirements. The information exchanged by doctors, researchers, patients, etc. in the MORMED platform may include personal health details which should not be made available in the cloud and put at Google's or anyone else's disposal. The environment needs to be closed and controlled in order to preserve the users' privacy.

As for Systran, it covers the language pairs English-German, English-Spanish and German-Spanish, in all directions. Its main features are:

- A rule-based translation system allows translating immediately without having to input a huge amount of parallel texts or dictionaries, thanks to the in-built dictionaries which come with the system. Besides this, Systran includes not only dictionaries for general translations, but also some dictionaries specialised in fields like Medicine, Life Sciences and Chemistry, among others, which are extremely useful for the project.

- The last released version, Systran 7, includes hybrid technology which combines Statistical Machine Translation capability along with Rule-Based Machine Translation. This will increase the work to be done on the linguistic resources, since not only will it be necessary to collect and prepare resources in the form of dictionaries, but also in the form of bilingual aligned texts. However, the amount of work is benefited by the resulting higher quality of the translation output.

- There is the possibility of controlling translation quality. The user can create dictionaries and include them in the system, and also force it to prioritise the user dictionary above system dictionaries. We thus ensure that terms related to Lupus and the Antiphospholipid Syndrome will result in a more accurate translation.

- The technology is mature, and has been used already for a long time in the European Union and with commercial customers across the world. Its translation output has proved to be understandable.

- It has a client–server application.

- It is easily integratable into the MORMED platform.

- The price of Systran is quite high and we invested in the latest enterprise version of this technology in order to use it as a training system prior to a later stage when ideally we switch to SMT (e.g. Moses) once the required volume of training data has been obtained.

Regarding the Hungarian language, it is difficult to find products on the market which offer combinations with this language.

The solution offered by MorphoLogic, called MorphoWord Pro, has been taken into consideration.

This is a purely RBMT system, also customisable by adding user dictionaries in the desired field or domain.

It is apparently not as powerful as Systran and looks slightly more rudimentary, but the evaluation tests carried out (see next section) show that it gives potentially usable translation results. Besides, it is the only sophisticated system available for Hungarian.

3.3. Evaluation, assessment of tools' possible features not considered in 3.1

3.3.1. Statistical vs. Rule-Based Machine Translation

Given the low volume of specialised data available in the domain of the MORMED project, no linguistic evaluation was done for Moses, since the results would not be reliable. The output generated by SMT systems improves dramatically with the amount of material on which they are trained, and the small size of the corpora available at the beginning of the project resulted in a decision against SMT at this stage.

As explained above, statistical machine translation makes sense in 2 main contexts:

1. when the domain is restricted
2. when large corpora are available to train the system.

This was confirmed during a test carried out with another SMT System (Language Weaver) evaluated against Systran prior to MORMED in the automotive domain. For test results please refer to chapter 4.

In our case, the domain may not be very broad. Within the medical domain, the diseases involved in the MORMED project are very specific. However, the limited amount of texts available regarding the subject matter means that we must disregard the use of statistical machine translation systems, at least in the initial phases of the project.

Therefore, the effort was focussed on the evaluation of the RBMT systems initially considered: Systran (for EN-ES-DE) and MorphoWord Pro (for EN-HU).

We evaluated Systran's functionality in detail in order to determine the required workflow for populating , maintaining and updating the system. The following issues were considered:

i) *Preparation of resources*.

Both for Systran and MorphoWord Pro some sample resources were prepared and fed into the system to check ease of use and translation quality. This involved:

· Feeding the translation system with a test dictionary and bilingual aligned texts as training data for the statistical component.

· Analysis of the translation output before and after populating the MT systems with customised resources (dictionaries and bilingual aligned texts), and comparison with human translation.

· Analysis of files translation: one document versus multiple documents at the same time in different formats.

The test resulted positive in both cases differently:

a. For Systran, it is possible indeed to add customised terminology to the machine translation system and obtain improved translation output. Given that Systran now includes hybrid translation technology, it is also possible to populate the system with bilingual aligned corpora as well. Therefore, the linguistic resources in Systran are in the form of dictionaries and aligned texts.

b. For MorphoWord Pro, it is possible as well to add customised terminology to the machine translation system and obtain improved translation output. However, this tool is purely rule-based. Therefore, the linguistic resources to be imported into this system will be only in the form of dictionaries. Besides this, we can take advantage of MorphoWord Pro's term extractor by extracting terms from specialised bilingual texts. MemoQ, a TM tool developed in Hungary, already has an interface with MorphoWord Pro, and will be included in the translation workflow proposed in deliverable D3.1 "MORMED Platform Architecture".

ii) *File formatting*. Both translation systems work with common file formats: .doc, .pdf, .tmx, .html, .xml, among others. If a user wants to have a document translated, it is important that the translation system provides appropriate filters. The output file format needs to correspond to the input format. Both for Systran and MorphoWord Pro input and output formats were checked and both of them support the most usual file formats.

iii) *Ease of use*. Although the machine translation system will appear as an internal layer to the MORMED platform and users will not access it directly, it still should be easy to use for a language expert who will be working in the background updating or adding new linguistic resources.

Some tests were done regarding this matter: creation of new resources, edition of existing resources, update of the term bases, translation tests, human post-edition of translation output, etc. They proved that it is indeed not complicated to manage the system.

iv) *Flexible and easy to control resources*. Some tests were also done regarding the flexibility of the resources. It is important that the resources, once inserted in the system, can be easily controlled and further customised. The test results were positive: it was indeed possible to manipulate the resources to obtain better translation output.

v) *Translation*. Several translation tests were carried out, and proved to result in usable output.

vi) *Populating and managing resources in the system*. The system will need regular updating of texts and dictionaries. It is important that this task can be done in a smooth and uncomplicated way. Tests showed that this is feasible in both translation systems.

vii) *Compliance with additional translation tools.* As stated in the Introduction to this deliverable, LTC Communicator II will require different translation tools interacting together. Therefore, it was checked whether both machine translation systems comply with TM tools such as memoQ and Trados, and with standard formats such as tmx, csv, etc. The results were positive.

viii) *Access*. It is important that access to the machine translation system is guaranteed through the MORMED platform. Remote access to the server –where the machine translation systems will be allocated– is done via internet through a secure connection. Both Systran and MorphoWord Pro are client-server applications which will be installed on the server and guarantee remote access.

ix) *Interface with LTC Communicator II*. Both Systran and MorphoWord Pro can and will be integrated with LTC Communicator II.

3.3.2. Systran vs. Language Weaver Evaluation Methodology

Test data

HTML files provided by Daimler – approx. 150 in total. We selected 7 files from the first batch as representative sample.

Procedure

We translated these files in Systran 5 (which at the time did not include an SMT component) with an automotive user dictionary generated from Trados TM and a list provided by Daimler – total 4,273 terms.

The same files were translated in the Language Weaver (LW) demo system populated with a bilingual corpus provided by Daimler (8 million words per language).

The source language was German and the target English.

4. Results of translation quality evaluation

4.1. Systran vs. Language Weaver

The Systran vs. Language Weaver automatic evaluation technique previously tested at LTC yielded the following results:

LW translations were superior to Systran when applied to a very narrow domain; most would be usable without post-editing (definitely not true of Systran).

Systran relies on user dictionaries (UD), which means a considerable human overhead to populate/maintain. Systran results were superior, though, when dealing with a broader domain.

Detailed comparison:

	Systran		Language Weaver
-	DE word order retained (e.g. object – verb)	+	EN word order (e.g. verb – object)
-	No capitals at start of segments	+	Intelligent control of upper/lower case
-	Unknown words not translated – left in DE	+	Nothing left untranslated
-	1:1 DE-EN (i.e. same EN word always used to translate a given DE word)	+	LW output more context-sensitive
-	Relies on comprehensive user dictionary, manually populated, ongoing maintenance	+	Builds own domain-specific vocabulary from analysis of corpus
-	Not intelligent with compounds – if word in UD used in compound, this also needs to be listed	+	Compounds derived from corpus
+	All DE words either translated or left in DE	-	Some words omitted (mostly verbs)
+	Whole file always processed	-	Some files truncated

Table 1 – Comparison of Systran and Language Weaver

We also used the *weighted N-gram model* (WNM) for comparative evaluation, which is an extension of the established *BLEU* method. (Babych & Harley, 2004)

An overview of the scores is shown in Table 2 below.

This proves that LW evaluation results are much superior to Systran based on training the system on a large corpus consisting of high quality relevant data in a very well defined narrow domain.

German	Reference translation	MT	Precision	Recall	F-score
a) Work instructions	Based on Daimler TM	Systran: LW:	28.4% 75.5%	31.0% 82.0%	29.6% 78.6%

Table 2 - Test carried out for translation German→English on sample texts. *Precision* indicates how closely the translation reproduces the contents of the reference text, and *recall* indicates the percentage of words in the reference that also occur in the translation –a rough measure of how much the meaning of the text has been conveyed.

In a second stage of the same evaluation cycle, we submitted Volkswagen texts to both systems, and here, the Systran results were better than the LW output. This means that translation quality deteriorates when the domain is less narrow or well defined.

This experience helped us determine the procedure to follow within MORMED.

4.2. Constraints

Within the MORMED project, the following constraints apply:

1. Systran's SMT component in the version we have available has a limit in the customised translation memories of 40,000 segments. Another Systran version with unlimited TM size exists, but we cannot afford to invest €150,000 in this highest level enterprise solution for the time being.

This limitation of the SMT component is not a problem for the duration of the project, though, since as mentioned above we are in the process of preparing the bilingual aligned corpora which at the moment have not exceeded this limitation. Once we have corpora available that exceed this limitation, we will export the data and retain them separately for use in the SMT system to be tested once sufficient resources have been translated and post-edited with Systran. We expect to use Systran as a training tool for ideally switching to SMT as soon as training data of several millions of words per language are available.

We will consider introducing a statistical machine translation system which can deal with a large amount of parallel text as soon as it is feasible. The technical and linguistic characteristics of the platform allow for it.

2. The internal process within Systran between its SMT and RBMT modules is not transparent. It is not possible for us to act on them, since we cannot know which parts of the document have been processed by which module. Systran appears to us as a black box which we can customise with User Dictionaries (for the RBMT module) and bilingual corpora (for the SMT module).

Despite this, we can take advantage of the SMT component by customising and feeding TM resources, as explained above.

4.3. Implementation process. Phases

After all tests were carried out, it was confirmed that both machine translation systems (Systran and MorphoWord Pro) are valid for the purposes of the MORMED project translation backend.

Both products match the selection criteria and translation needs of the MORMED project to a large extent.

Therefore, we can confirm the implementation feasibility of the process as shown in Figure 5 on page 27.

In a first phase, only a Rule-Based Machine Translation system will be used, namely Systran. We will customise Systran and use it as a training system, by creating domain specific user dictionaries and. bilingual corpora feeding Systran's SMT component, as well as the TM and SMT systems in phases 2 and 3.

A second phase may follow where a TM module is included as a translation step before the RBMT system. In this phase, the translation output from the TM will be limited to 100% matches. This means that only fully matched translations from the TM system will continue on to the platform, whereas those segments which could not be translated 100% by the TM system will be automatically passed on to the RBMT system. The machine translated segments will be flagged for human post-editing and the high quality result will be fed into the TM automatically after post-edition.

It was decided to limit the TM output to 100% matches, as the full translation process is in batch mode and contrary to some other opinions in our industry, we believe that proposing alternatives to human post-editors, including lower level fuzzy TM matches, will be unpractical.

In a third phase, the RBMT system is replaced by a SMT system (Moses), once we have collected a sufficient amount of bilingual aligned corpora. Tests with Moses will start once a 1 million words threshold per corpus per language is reached. If Moses is not appropriate, we will consider Language Weaver, but have no budget to include an interface to a commercial SMT system in our overall solution.

The new LTC Communicator tool will continue to be developed alongside and according to these phases, and will contain MORMED specific features.

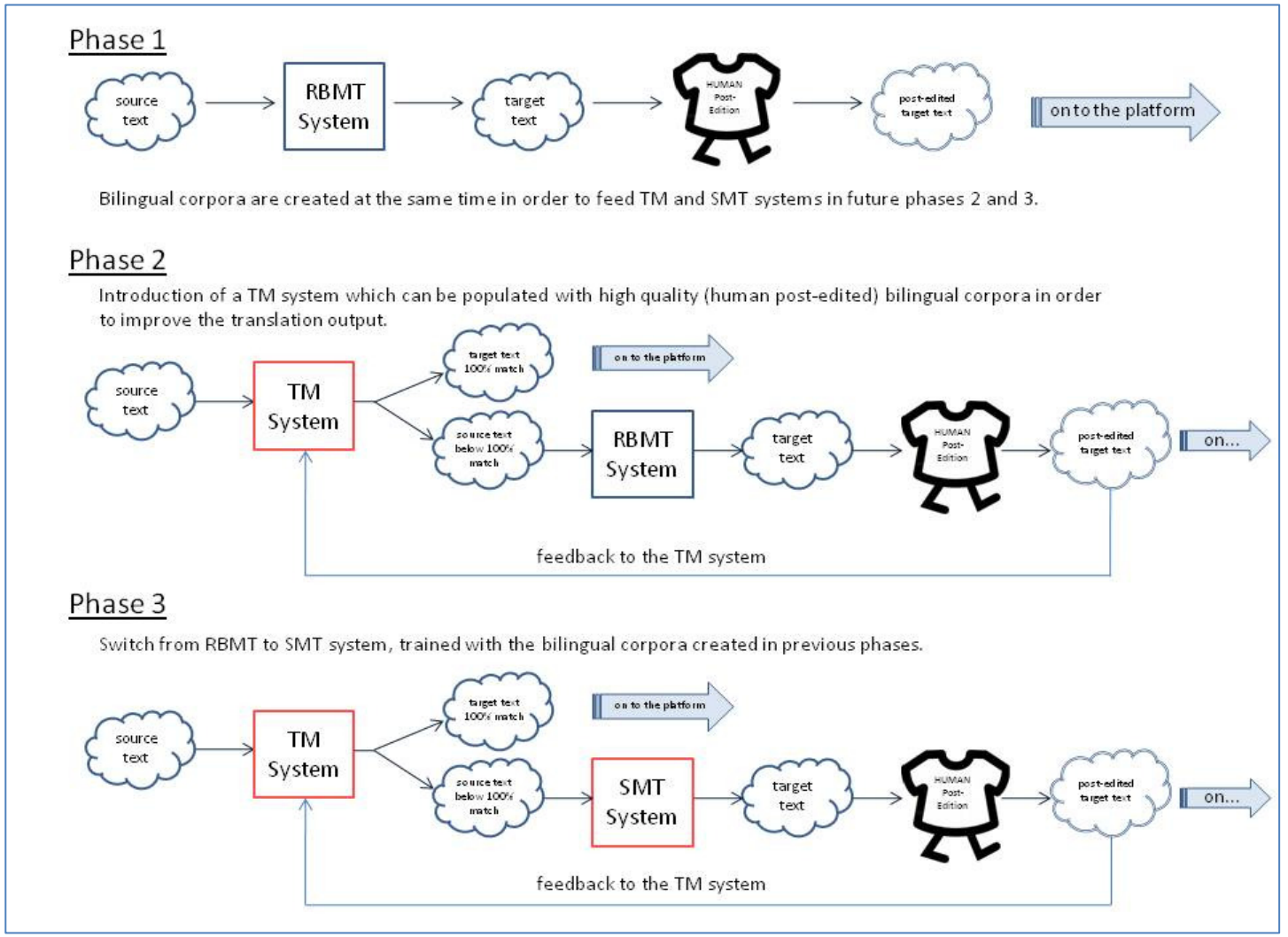

Figure 5 - Linguistic workflows and phases

5. Conclusions

The aim of this document was to describe evaluation techniques and results of different machine translation tools in order to integrate the most suitable one(s) into the MORMED platform.

An overview of the current machine translation technology was given. There are mainly three types of machine translation systems on the market: statistical machine translation, rule-based machine translation and example-based machine translation (translation memory technology).

Also, the criteria and requirements for the translation backend of the MORMED platform were defined, and according to these several machine translation technologies were considered.

In a first instance, Moses and Language Weaver (Statistical Machine Translation systems) were considered. They had to be disregarded due to an initial lack of appropriate linguistic resources to train the system, as statistical machine translation needs a lot of training data (in a narrow domain) in order to yield acceptable and understandable translation output.

At a second stage, Systran (an initially Rule-Based, now hybrid Machine Translation system) was evaluated. The evaluation was satisfactory for the MORMED platform needs, but insufficient for the languages considered in the project (English, German, Spanish and Hungarian). It supports the first three ones, but another translation system would be needed for Hungarian.

Thirdly, MorphoWord Pro (a Rule-Based Machine Translation system) was evaluated for the Hungarian-English language pair. The tests were satisfactory as well.

As a result of this work, initially two machine translation systems will be integrated into the LTC Communicator supporting the MORMED platform: Systran (for EN-DE-ES) and MorphoWord Pro (for EN-HU). At a later stage, once the system has more bilingual resources available, a switch to statistical machine translation is foreseen. We hope to be able to use Moses, as no funds have been foreseen to buy Language Weaver, which was recently acquired by SDL

Bibliography and References

Babych, B., & Hartley, A. (2004). *Extending the BLEU MT Evaluation Method with Frequency Weighting*. University of Leeds, UK.

Babych, B. (2004). *Weighted N-gram model for evaluating Machine Translation output*. University of Leeds, UK.

Cameron Shaw Fordyce, Xavier Gros. (2007). *Survey of Machine Translation Evaluation.* Saarbrücken: EuroMatrix.

Cocci, L. (2007). CAT Tools: Istruzioni per l'uso. *Daf Werkstatt* , pp. 133-147.

Hutchins, J. (2009). *Compendium of Translation Software.* European Association for Machine Translation. http://www.eamt.org/soft_comp.php (last date accessed November 2010).

Hutchins, J. (2005). Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation, vol. 17* , pp. 5-38.

Mason, J., & Rinsche, A. (1995). *Translation Technology Products.* London: Ovum Ltd.

Papineni, K., Roukos, S., Ward, T., & Wei-Jing Zhu (2001). *BLEU: a Method for Automatic Evaluation of Machine Translation.* IBM T. J. Watson Research Center, USA.

Puntikov, N. (1999). *MT and TM technologies in localization industry: the challenge of integration*. Machine Translation Summit VII. Kent Ridge Digital Labs. Singapore.