Project acronym: LISE

Grant Agreement Number: 270917

Project title: Legal Languages Interoperability Services

Final LISE Service – Version Three

D2.2.3 – Final LISE Service version

Dissemination Level: PU

Version No. Final

22/02/2013

Project co-funded by the European Commission within the ICT Policy Support Programme.

1. Document Information

Deliverable Number	D2.2.3 – Third LISE Service version (Draft)
Deliverable title	LISE Service Version Three Short Report
Due date of deliverable according to DOW	31 January 2013; extended to 22nd February 2013
Actual submission date of deliverable	22nd February 2013
Main Author(s)	Michael Wetzel, ESTeam; Lambros Kranias, ESTeam
Participants	Joeri van de Walle, CrossLang; Gudrun Magnusdóttir, ESTeam; Markos Xagoraris, ESTeam
Reviewer	Tanja Wissik, UniVie
Work package	2
Work package leader	ESTeam
Dissemination level	PU
Version	1.3

Revision History

Revision	Date	Author	Organisation	Description
1.0	**25 January 2013**	**Michael Wetzel**	**ESTeam**	**1st complete draft**
1.1	**31 January 2013**	**Michael Wetzel**	**ESTeam**	**Incorporated feedback from ESTeam**
1.2	**04 February 2013**	**Michael Wetzel**	**ESTeam**	**Document text review**
1.3	**18 February 2013**	**Michael Wetzel**	**ESTeam**	**Incorporating Feedback from Luxemburg/IATE workshop, Ghent EMB/TMB meeting and Ghent Advisory Group meeting**
Final	**21 Feb. 13**	**Gudrun Magnusdóttir**	**ESTeam**	
	22.Feb.13	**Tanja Wissik**	**UniVie**	**Formal Review**

Statement of originality:
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

2. Table of Contents

1. Document Information 2
3. Executive Summary 4
4. Data Processing 5
LISE Domain Data Selection 5
Meta Data Customization and Data Processing 5
ESTeam Tools Changes 6
5. Statistics and Processing 8
IATE Data and LISE Subset 8
Cleanup Processing Results 9
OMEO Processing Results 10
FILLUP Processing Results 12
6. Input from the User Community 13
7. Conclusion 15

3. Executive Summary

In version one (M6) the main portal and infrastructure for the LISE Web Service and the Human Support Interface was delivered. The service is since then continuously available at https://app.lise-termservices.eu.

Version two (M12) of the web service was significantly improved. The service was equipped with a fully featured collaboration functionality. Site members could start new discussions (so-called *topics*), reply and comment on those. See D2.2.5 Human Support Interface for a detailed summary.

For the past 12 months work has focused on providing advanced data processing and including this in the deliverables of LISE (LISE Version 3 - M24). IATE, perhaps the biggest terminology database available, has been analyzed, restructured and processed to enable work on a suitable subset as set forth in the LISE proposal. Relevant data categories (Meta data) have been identified and specific rules for processing have been implemented.

Due to the scope of the information and classification available in IATE, ESTeam Language Server has gone through significant changes to process the data, including the invention of new algorithms and modification of previous existing algorithms. Accordingly, all three ESTeam Tools have been changed and enhanced to become as general purpose as possible. However, in order to process terminology in this advanced way it has been concluded that the tools must be specifically customized to be efficient due to the variation in which terminology databases are set up. The tools are available to the LISE Consortium and to the IATE user group.

User feedback for the LISE Collaboration Portal as well as the ESTeam Tools has been taken into consideration and crucial changes have been implemented.

4. Data Processing

As planned and committed in D2.2.1 and D.2.2.2, this LISE iteration (M12-M24) is focused on data and data processing. This involved identification and selection of specific data, cleaning data management and uploading to all environments to be able to undergo the processing required by each tool.

This meant introducing a large number of changes to the ESTeam Language Server and ESTeam Tools, which were originally designed for IP classifications, but are now optimized for processing general purpose terminology databases of multiple type and origin such as is the case of the IATE data.

The result of the processing has been proven to be extremely successful in all three cases but for the users perhaps especially for the two tools Cleanup and Fillup (please see the Evaluation Report).

LISE Domain Data Selection

Together with WP1, WP3, and WP4 participants, ESTeam has analyzed and then identified several domains (domain attribute value) in the IATE data. This was preceded by several analysis runs that ESTeam presented at the LISE EMB/TMB meeting in Reykjavik (July 2012). This led to identifying concrete domain values to focus on; it is noteworthy that the LISE Consortium used the LISE Collaboration Portal / Human Support Interface for discussing the data selection.

In order to be able to use the tool Fillup it was also necessary to identify and qualify appropriate translation memory (TM) data to support the terminology data. The first proposal was to use the TM resource of the Acquis Communitaire but that turned out to be useless – since the domains / topics did not match this provided no added value. The IATE group then kindly submitted another translation memory resource which matched better and this was also enhanced by the Austrian Parliament who provided further data.

All of the TM data underwent further data processing including sub-sentence alignment and uploading with metadata into ESTeam databases, to support the Fillup tool which works on assigning translation suggestions to small units such as is often the case with terms.

Meta Data Customization and Data Processing

ESTeam development made several changes to the ESTeam Language Server, where the main processing of the data takes place: adapting import formats, interpreting several Meta data fields that are proprietary to IATE, tuning algorithms.

The software changes have led to an analysis which can clean, improve, and enhance the terminology resource. Intermediate results have been discussed in several meetings and the final results presented and discussed with the IATE group at a joint workshop with all LISE

partners on 5th February 2013. In the WP4 report there will be a detailed qualitative and quantitative analysis of the user perspective and workshop outcome.

ESTeam Tools Changes

The ESTeam Tools went through several feature and user interface changes. The most important are listed below:

Tool and Focus	Software Changes

ESTeam Cleanup	• Elaboration on each of the Cleanup Rules
	  ○ Spelling: Enhanced reference dictionary to better cover the LISE domain
	  ○ Canonization: Enhanced reference dictionary to better cover the LISE domain
	  ○ Language: Tuning of the decision algorithm
	  ○ Equivalent: Enhancement to allow for linguistic similarity (and not just string-identical terms)
	  ○ Domain: Adaptation to LISE domain
	  ○ Translation: Tuning of the decision algorithm. Adding a new set of results derived from the new OMEO multilingual linking process (see below)
	  ○ Subset: Enhancement to also allow for linguistic similarity (and not just string-identical terms). It now also locates language-specific subsets (so-called *partial subsets*)
	• User Interface Rework:
	  ○ Re-designed parts of the software (CleanUp Admin application & CleanUp User application)
	  ○ Introduced several new features: direct link to IATE entry, search, evaluation reporting, results exporting, “dual mode” for the Translation error category
	• Continuous and iterative accommodation and processing of the domain data

ESTeam OMEO	• User Interface Rework:
	  ○ Re-designed parts of the software applications (Omeo Admin application & Omeo User application)
	  ○ Introduced new features: direct link to IATE entry, exporting of results
	• Invented and implemented the *Multilingual Linking* process which results in broader IATE groups (for the purpose of Fillup) and a set of potential Translation category errors for Cleanup
	• Continuous and iterative accommodation and processing of the domain data

ESTeam Fillup	• Triaged, gathered and processed (sub-sentence alignment) available translation memory data (limited) • Created the reference translation memory • Tuned the Fillup processing (types of processing, thresholds)

- User Interface Rework
  - Re-designed parts of the software applications (FillUp Admin application & FillUp User application)
  - Introduced new features: Completion mode, exporting of results exporting
- Continuous and iterative accommodation and processing of the domain data

5. Statistics and Processing

LISE WP2 focused in this iteration first on IATE as a whole and then on a subset of IATE selected to be processed in detail in LISE, as well as demonstrated to the IATE users.

IATE Data and LISE Subset

ESTeam Language Server processed in total 1,470,943 IATE entries, covering up to 22 languages, spanning 11,163473 terms. The terms however are not distributed equally across languages. English and French see a wider distribution than the "newer" EU languages like Latvian or Bulgarian:

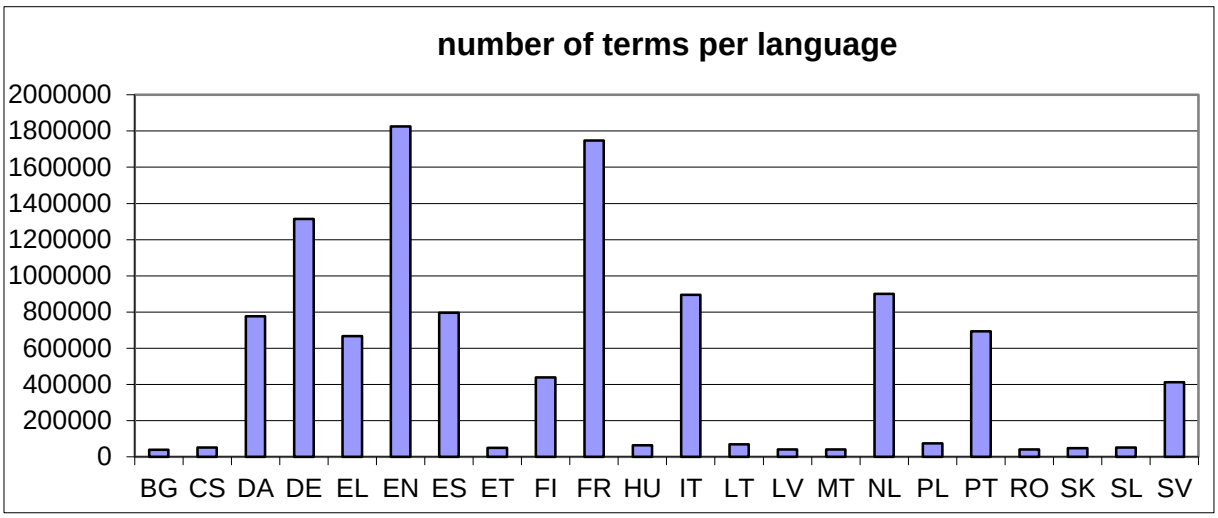

Figure 1 – Terms / language distribution in overall received IATE data export

After having selected the data for LISE, namely entries that cover administrative, legal, social security terms, the following distribution of this subset of the IATE terminology data can be seen: in total 21,515 entries, comprising 95,544 terms.

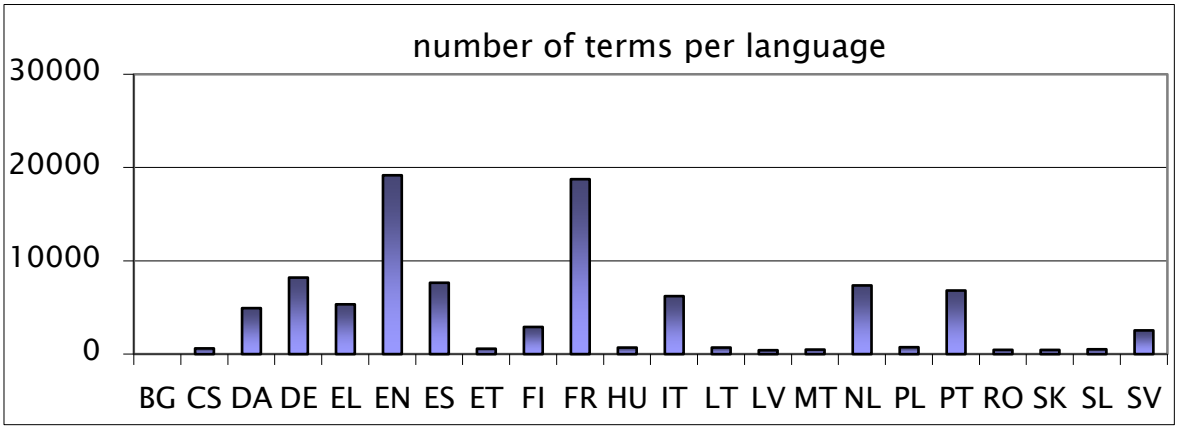

Figure 2 – Terms / language distribution in LISE relevant IATE data

For the ESTeam Fillup process it is also interesting to see how many entries cover how many languages. For the LISE relevant data (21515 entries), it was analyzed that 8758 entries (~41%) contain only one language, i.e. are monolingual entries; 4375 contain two languages, 2806 contain three languages, and only very few entries cover all 21 languages (85 entries, i.e. ~0.4%). This indicates the need for a half automated, ESTeam Fillup supported process to enhance the IATE data.

The analysis of the concrete outcome of the ESTeam Tools will be half quantitative and half qualitative. Availability of the quantitative figures will be synchronized with WP4. Availability of the qualitative figures, i.e. human estimation on how good the outcome is, will be incorporated here after the next workshop with the IATE user group.

Cleanup Processing Results

Below matrix lists the amount of spotted errors per category. For instance, for EN (English), ESTeam Cleanup spotted 54 potential translation errors in the IATE data (LISE domain).

	TOTAL	SPELLING	CANONISATION	LANGUAGE	EQUIVALENT	DOMAIN	TRANSLATION	SUBSET
EN	310	22	0	4	6	225	54	102
FR	387	24	7	25	10	282	42	154
DE	121	14	0	7	7	68	25	67
ES	170	22	4	11	8	118	8	28
IT	158	14	0	12	1	122	9	33
DA	81	9	0	28	2	38	4	19
NL	113	14	0	20	3	63	13	33
SV	40	6	0	13	0	16	5	8
EL	124	7	1	0	2	101	15	18
PT	119	8	2	9	3	95	2	27
FI	28	9	0	8	0	8	3	7
CS	17	5	0	6	0	5	1	1
HU	14	7	0	4	0	2	1	1
PL	14	4	0	3	0	4	3	0
SK	13	6	0	2	0	5	0	1
SL	6	3	0	1	0	2	0	2
ET	10	8	0	1	1	0	0	1
LT	12	3	0	4	0	4	1	2
LV	10	6	0	2	0	2	0	1
MT	7	0	0	4	0	2	1	0
BG	0	0	0	0	0	0	0	0
RO	8	1	0	3	0	2	2	1

The results were displayed in the Cleanup tool fully linked to the IATE database since this was requested in a first meeting with the users. An example can be seen below:

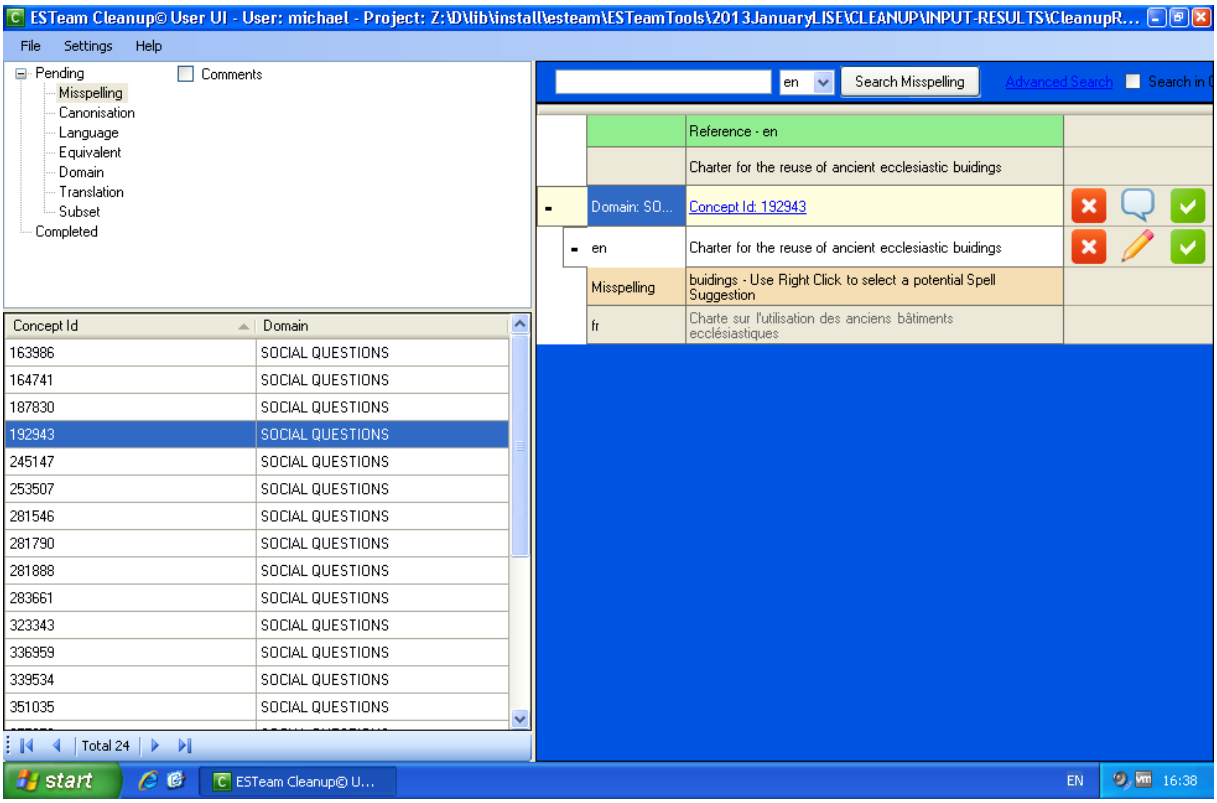

Figure 3 – ESTeam Cleanup, Category 'Misspelling': Spotted *buidings* in English term; Direct link to IATE id #192943.

OMEO Processing Results

During the first round (*monolingual grouping*) ESTeam OMEO identifies identical IATE entries by comparing similarity of terms in the same language. The top five languages were: French (904 entries), English (766), German (476), Dutch (347), and Italian (326). The outcome is then OMEO groups, spanning the terms from several IATE entries.

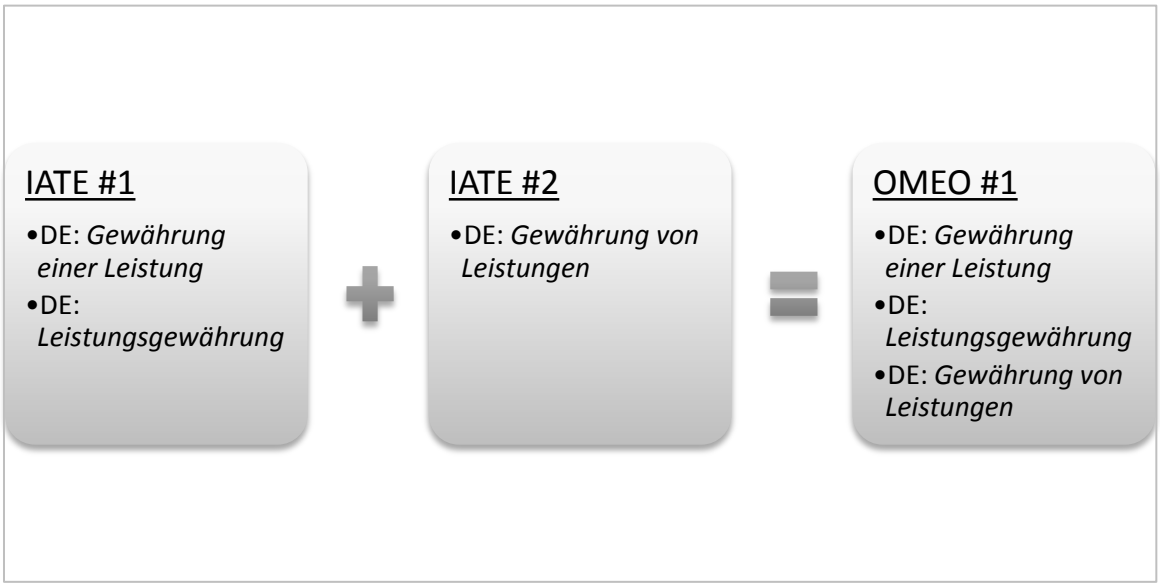

Figure 4 – OMEO First Round, Monolingual Grouping

Figure 4 explains how the German terms are taken to build a new group, OMEO #1.

Now, combining the results of OMEO #1 again with the original, multilingual entries IATE #1 and IATE #2 creates an even bigger group; terms of many languages expressing all the same concept, see Figure 5.

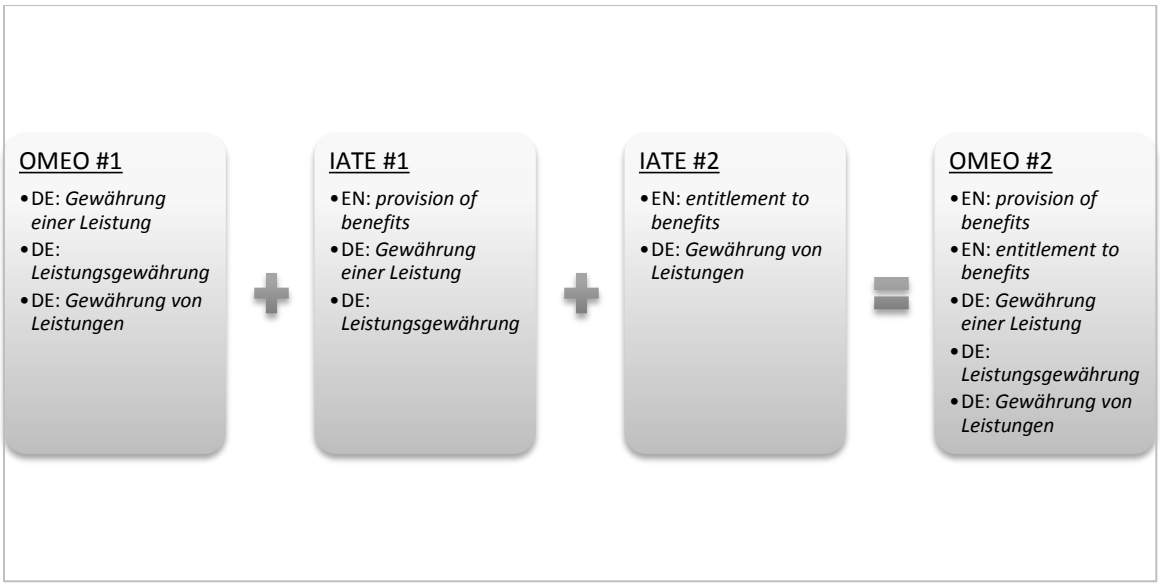

Figure 5 - OMEO Second Round, Multilingual Grouping

In this the second round the outcome of the first round, the broader OMEO groups are therefore playing a pivot role – and are then combined with the IATE multilingual entries. This process then created 19684 multilingual groups (like OMEO #2). In Figure 6, we see that English or French are now present in ~12,000 groups, German in ~5,000 groups etc.

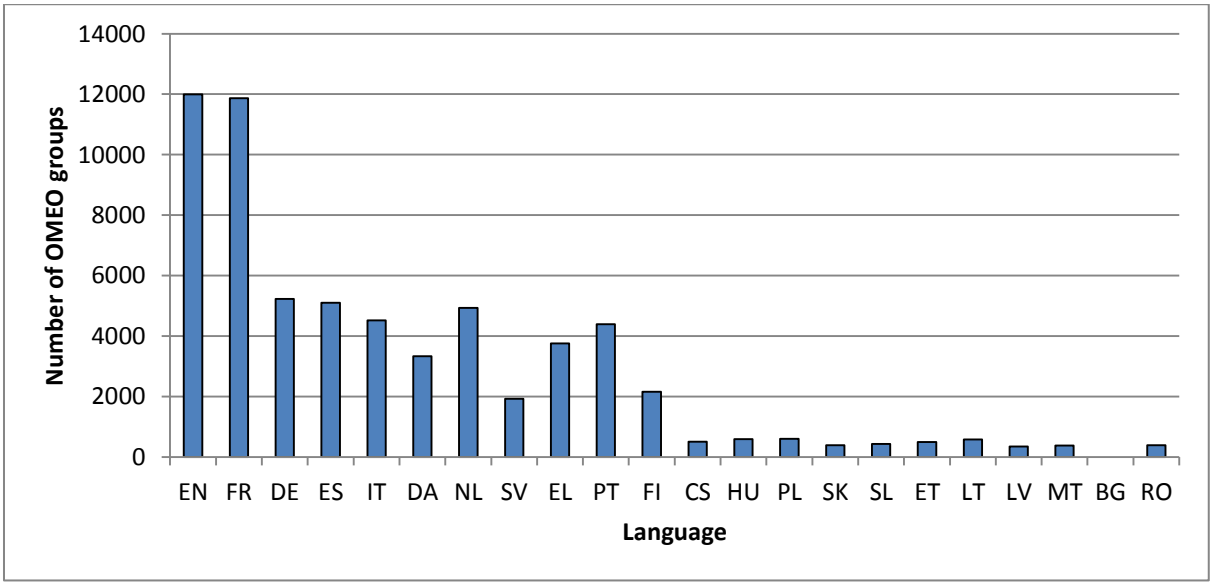

Figure 6 – OMEO Multilingual Grouping: Languages represented / amount of groups

Making these groups as broad as possible makes the ESTeam Fillup process more efficient, since wider translation suggestions can be applied.

FILLUP Processing Results

While the available TMs were not perfectly matching the LISE domain, ESTeam Fillup still comes up with noteworthy suggestions, proving its general applicability to the use case. It also shows that the outcome of ESTeam Fillup obviously depends on the applicability of the translation memory.

Translation Direction	# of groups to enhance	# of groups with a suggestion
EN – DE	4280	356
EN – FR	3121	228
EN – ES	4187	375
EN – EL	4622	366
DE – EN	606	45
FR – EN	2492	258
ES – EN	755	50
EL – EN	299	14

Details about these figures – as well as on ESTeam Cleanup and ESTeam OMEO – and how to interpret them can be found in the WP4 deliverable, the D4.2 Evaluation Plan. The results were displayed to the users and presented in the IATE workshop as can be seen below:

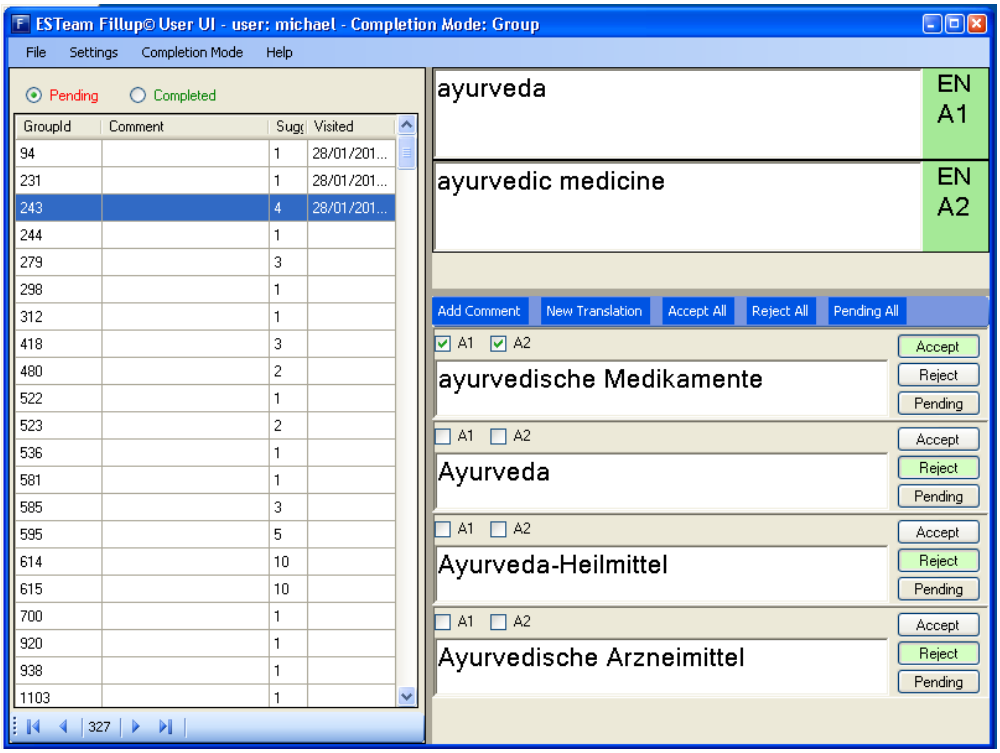

Figure 7 – ESTeam Fillup: Harvested TM to suggest several translations for ayurveda / ayurveda medicine

6. Input from the User Community

Even if a lot of the work in the third version took place in the ESTeam development labs, there have been significant interaction and feedback from users both within LISE and the IATE group.

LISE partners and the IATE group helped to identify the relevant IATE data including translation memory resources. A first IATE/LISE meeting was held in Luxembourg in October 2012 and was successfully attended by all relevant stakeholders in IATE, including the EU Commission, Parliament, Council and the CdT together with LISE partners, ESTeam and CrossLang.

The ESTeam Tools capabilities and status were demonstrated and discussed with approximately 15 participants. A very important feedback from IATE users was that users prefer to directly edit the entries in the IATE database – and not in the ESTeam Tools, leading to the implementation of the direct linking feature in the Tools (see Figure 3). Further, a request for a "memory" functionality, i.e. tracking capabilities, was proposed – so that the ESTeam Tools do not re-list already rejected suggestions.

A second joint workshop between the IATE group and the LISE project partners took place on 5th of February. A preparatory, several hours online session was held on 23rd January 2013, with LISE project partners that led to several good suggestions on how to improve the tools and the processing. Many of the suggestions, for instance monitoring changes so that evaluation is more transparent were implemented before the workshop with the IATE group.

Figure 8 – IATE Workshop 5th Feb 2013, Luxembourg

Detailed results of the workshop are described in the WP4 Evaluation Plan deliverable. However, we would like to mention here that for instance feedback like "this is a

terminologists's dream" when referring to ESTeam Cleanup is obviously very encouraging. More workshops with IATE users are planned, underpinning the usefulness of language technology processing in the area of terminology within the EU institutions. A second workshop is scheduled in Brussels and being planned upon writing this document.

7. Conclusion

The LISE project proved how useful and applicable the ESTeam Tools are for terminological work outside of the trademark domain. The ESTeam Language Server and all Tools have been changed to be more generic but still customized to store and process term entries from IATE.

Even if progressing toward a facilitated customization effort for the ESTeam tools, an important learning is the fact that making general purpose software out of the ESTeam Tools is not viable. The tools must be seen as a high end customization and service offering. In all cases, the individual needs, systems, and workflows are too different, so that customization will always be required; see also D6.2 Exploitation Plan for further discussion regarding this point.

For the remainder of the LISE project, ESTeam foresees an on-going data processing effort to support the workshops in a valuable way, as well as supporting changes based on user feedback respectively WP4 evaluation.