FP7-ICT-2011.4.2
Contract no.: 288342
www.xlike.org

Deliverable D7.3.2

Second Benchmarking Report

Editor:	Marko Tadić, UZG
Author(s):	Marko Tadić (UZG), Božo Bekavac (UZG), Matea Srebačić (UZG), Daša Berović (UZG), Danijela Merkler (UZG), Tin Pavelić (UZG), Achim Rettinger (KIT), Lei Zhang (KIT), Xavier Carreras (UPC)
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	M24
Actual Delivery Date:	M24
Suggested Readers:	All partners of the XLike project consortium and end-users
Version:	1.0
Previous Versions:	
Keywords:	evaluation, linguistic analysis, natural language processing, named entity recognition and classification, semantic annotation, knowledge extraction

Disclaimer

This document contains material, which is the copyright of certain XLike consortium parties, and may not be reproduced or copied without permission.

All XLike consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the XLike consortium as a whole, nor a certain party of the XLike consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	Cross-lingual Knowledge Extraction
Short Project Title:	XLike
Number and Title of Work package:	WP7 – Multilingual Linguistic Processing
Document Title:	D7.3.2 – Second Benchmarking Report
Editor (Name, Affiliation)	Marko Tadić, UZG
Work package Leader (Name, affiliation)	Pat Moore, Bloomberg
Estimation of PM spent on the deliverable:	12 PM

Executive Summary

This document gives a report on the evaluation of different processing methods developed during the Y2 of the project. The methods developed belong primarily to WPs that deal with the preprocessing stages of the general XLike pipeline, namely, the linguistic preprocessing (WP2) and conceptual mapping (WP3). This document is the second of three (T7.3.1 Y1, T7.3.2 Y2, and T7.3.3 Y3) that are associated with benchmarking the methods developed within XLike. It also refers to the B1.1.3 Indicators and Metrics part of the DoW where expected target outcomes for different categories are defined and respective progress tracked.

Specifically we have developed a RECSA, Resource for Evaluation of Cross-lingual Semantic Annotation, a parallel corpus manually annotated, that is used for evaluation of methods developed in WP2 and WP3, but it was also submitted to the LREC2014 conference to be available to wider Language Technology and Knowledge Technology research communities.

Here we give results of evaluation tests for (1) WP2: Named Entity Recognition and Classification modules for three XLike languages (en, es, de), and (2) WP3: the Final Text Annotation Prototype, a multi-lingual text annotation tool with a cross-lingual knowledge base, namely Wikipedia, for three XLike languages (en, es, de).

The performance of our implementations, evaluated with the new RECSA resource, that represents a scenario closer to a real life situation, is below the state-of-the-art. During year 3 we will analyze the causes for this, and make improvements to meet state-of-the-art accuracies.

Table of Contents

List of Figures

List of Tables

Abbreviations

CoNLL	Conference on Computational Natural Language Learning (http://ifarm.nl/signll/conll)
NE	Named Entity
NLP	Natural Language Processing
PoS	Part of Speech tag
RECSA	Resource for Evaluation Cross-lingual Semantic Annotation
SL	Source Language
TL	Target Language
NL	Natural Language
FL	Formal Language
MT	Machine Translation
SMT	Statistical Machine Translation
LT	Language Technologies
KT	Knowledge Technologies

Definitions

Parallel Corpus
Parallel corpus consists of documents that are translated directly into different languages.

Comparable Corpus
Comparable corpus, unlike parallel corpora, contains no direct translations. Overall they may address the same topic and domain, but can differ significantly in length, detail and style.

Source language
Language of the text that is being translated.

Target Language
Language of the text into which the translation is being done.

Formal language
Artificial language that uses formally defined syntax.

Language pair
Unidirectional translation from the SL to TL. Translation from L_a to L_b is one language pair and from L_b to L_a is another language pair.

Pipeline
Refers to the flux of different processes that are applied to a set of raw data in order to analyze it and interpret it. In NLP, a pipeline is a process that receives raw text and computes linguistic analysis, by a series of processes that perform morphological, syntactic and semantic analysis.

Let'sMT!
A platform for building, maintaining and using statistical machine translation systems out of your own data. This platform is the final outcome of ICT-PSP project Let'sMT![1] and it is available for registered users.

CycL
Cyc Language is an ontology language closely connected to Cyc ontology which in turn is the part of Linked Open Data. CycL is the formal language used for representing knowledge in Cyc ontology and it is defined as a declarative language based on classical first-order logic (relationships), with additional modal operators and elaborated quantifiers.

[1] http://www.letsmt.eu

1 Introduction

The benchmarking in XLike project is planned by DoW (section B1.1.3) in order to check whether the developed methods and tools, and by the end of the project the XLike pipeline in general, perform as expected. For different methods different evaluation scenarios are foreseen, but in general it can be said that we expect the performance near the state-of-the-art as reported in the referrent literature.

In the first benchmarking report (D7.3.1) we covered the evaluation of the methods and tools developed in Y1 of the project. These methods belong primarily to WPs that deal with the preprocessing stages of the general XLike pipeline, namely, the linguistic preprocessing (WP2, T2.1.1 and T2.2.1) and early prototype conceptual mapping (WP3, T3.1.1). Specifically, for WP2 we gave results of benchmark tests for PoS-tagging, lemmatisation, named entity detection and dependency parsing for six XLike languages (en, es, de, ca, sl, zh), while for WP3 we presented the performance of shallow multi-lingual text annotation tools with a cross-lingual knowledge base for three XLike languages (en, es, de).

In this second benchmarking report we cover the evaluation of the enhancements to the existing methods or to the methods additionally developed in Y2 of the project. These methods also belong to WP2 and WP3, i.e. the linguistic preprocessing (NE detection in particular) and conceptual mapping (using Wikipedia and Statistical Machine Translation techniques) serving as semantic annotation.

During Y2 the internal format, i.e. XLike XML schema was expanded in order to accommodate new XML elements and encode relations between them. We had to adapt our evaluation tools and develop additional automatic and manual evaluation scenarios. New schema features include new elements `<nodes>`/`<node>` and `<annotations>`/`<annotation>` that encode Named Entities and individual words that are linked to a conceptual spaces (Wikipedia, Wordnet, SUMO, OpenCYC, etc.) in a stand-off manner (more on new XLike XML schema see in D2.2.2).

Since after the evaluation presented in D7.3.1 we stated "Evaluation of methods used in early prototypes developed within WP2 show that the performance of our implementations is slightly below the state-of-the-art.", we wanted to make a more thorough check of the key methods used in the whole XLike processing pipeline in order to avoid the propagation of errors to the further stages of processing. In this context NEs are seen as one of the most important "information carriers" in news streams since they directly introduce entities from non-textual universe into the universe of the discourse. Their "information or knowledge load" is high, so the quality of this module can be seen as one of the most important in the whole process. Although the methodology for evaluation of **Named Entity Recognition and Classification** used and presented in the first evaluation report (D7.3.1) was already well-known CoNLL 2003 Shared Task [TM03], we were not satisfied by the results of our processing tools. We wanted to have a more detailed insight into the typology of errors within our results.

Additionally, for detecting whether the "lexical groundings" for concepts were recognized properly, i.e. whether the semantic annotation of texts by links to different conceptual spaces were established in a proper manner, we had to develop a new type of **Golden Standard** resource since for this type of evaluation none existed so far. The preliminary evaluation of Early Text Annotation Prototype (D7.3.1), although provided satisfactory results, was excercised only on a limited set of legal documents from the parallel JRC-Acquis corpus [STE2006]. This set was used at that time as out-of-the-box solution for a parallel corpus in order to provide a proof of concept for Early Text Annotation Prototype and to check whether this direction of research would give acceptable results at all. However, the genre of texts (legal documents) used for that previous evaluation campaign was far from the intended genre to be processed by XLike pipelines (predominantly news, but also social media contributions).

This is why we opted for a new evaluation scenario and why we decided to build the first Golden Standard for evaluating cross-lingual semantic annotation (to the best of our knowledge).

2 Building a Golden Standard for evaluating cross-lingual semantic annotation

The knowledge extraction from text can utilize the facts from e.g. DBpedia [BIZ2009], Freebase [BO2008], or Yago [HOF2013] as seed knowledge for the discovery of the relevant extraction patterns in large volumes of texts [KRA2012]. On the other hand these technologies can help to grow the knowledge base by automatic extraction of knowledge from text documents. From the side of language technologies, wordnets [FEL1998] or automatic ontology population methods [BUI2008] represent similar resources and techniques.

At the core of such technologies is the ability to relate words and phrases in natural language texts to existing resources in a knowledge base. If successful, a semantically annotated text document allows automatic contextualization and inference about the content of the document. Obviously, this task is highly language dependent, both on the side of the text document and the specific language interface to the knowledge base. In order to connect information across languages, efforts have been made on two levels:

1) machine translation systems can connect multilingual text documents to each other,
2) multilingual KB resources have been linked across languages (e.g. through language links in Wikipedia) or have been lifted to a language independent representation (e.g. Wikidata).

By combining techniques from both levels, the ultimate goal should be to construct cross-lingual semantic annotation tools that can link words and phrases in one language to structured knowledge in any other language or to a language independent representation. This is precisely what all tasks in WP3 are trying to achieve.

There have been extensive analyses of each of the tasks separately:

a) For machine translation evaluation efforts, see for instance [PAP2002]; [HAN2012],
b) Semantic annotation evaluation efforts [McN2009]; TAC_KBP[2] in 2013 address multilingual entity linking but not cross-lingual linking), and
c) Wikipedia cross-language links analysis [MEL2010]; [RIN2012].

However, there have not been any attempts in evaluating cross-lingual semantic annotation tools as a whole. This is why we developed a resource that can be used as a golden standard, i.e. a standard test set for evaluating and benchmarking cross-lingual semantic annotation systems collected from real life data. This resource was necessary to complete the task T7.3 as it was foreseen in DoW (p24), but in the same time the description of the work on building this resource was also submitted as a paper for LREC2014 conference.

This **Resource for Evaluating Cross-lingual Semantic Annotation (RECSA)** is composed of 300 news articles in three different languages (English, German and Spanish) with 100 articles in each language. The source of texts is a non-profit community of authors and translators Global Voices portal[3], that bring news reports in 35 different languages. The contens of this portal is available for use under CC-BY license, so the IPR status of texts is cleared by providing the reference to the online source. This opportunity opens this Golden Standard for many free future uses within the Language Technologies and Knowledge Technologies communities.

All 100 articles in this resource run in parallel, forming a trilingual sentence-aligned parallel corpus, thus allowing the investigation of cross-lingual semantic annotation techniques that need to keep the multilingual content under control, i.e. the content is the same in all three languages. The articles were

[2] http://www.nist.gov/tac/2013/KBP/data.html
[3] http://www.globalvoicesonline.org

downloaded, their boiler-plates were removed and they were converted into a plain text for further processing.

	English	German	Spanish
Tokens in total	74 337	79 146	77 476
Tokens per article	743.37	791.46	774.76
Sentences per article	27.12	34.62	27.36

Table 1. Basic statistics on RECSA resource

The desired result of producing RECSA at the first layer was to have a resource that will have:

1) Named Entities annotated and classified;
2) general concepts mentioned in text also annotated.

In the next step both of these "lexical groundings" are then linked to their respective Wikipedia articles. Having all this information annotated in a trilingual parallel corpus, provides the opportunity to measure the quality of the systems that try to establish links between texts and Wikipedia articles in a monolingual and cross-lingual context.

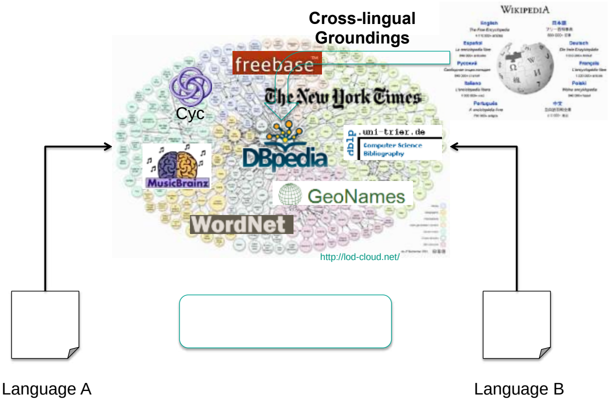


Figure 1. Cross-Lingual Semantic Annotation using DBpedia

English Wikipedia was selected as the most encompassing conceptual resource that is explicitly cross-linked to many languages, i.e. expressed in many languages. For this reason English Wikipedia was used as a hub conceptual space that also exhibits direct one-to-one links to DBpedia, which again, is linked with WordNet and numerous other linked data sources (see Figure 1). The lexical groundings in German and Spanish documents are linked to English Wikipedia also.

The first step in producing the RECSA was to get the parallel corpus annotated for NEs including their categories (Location, Person, Organization, Miscellaneous) and for general concepts.

The second step consisted of adding additional annotation with links of NEs and general concepts to Wikipedia.

In order to speed up the manual annotation, we first run the texts through the WP2 linguistic processing pipelines (see D2.2.2) for English, German and Spanish in order to receive automatic stand-off NE annotation. This annotation was then manually verified and cleaned, so that we can get the clear resource for NE layer.

The following step was the application of a semantic annotation method developed in the T3.1 (see D3.1.2), based on a newly developed cross-lingual linked data lexica, called xLiD-Lexica[4]. The results were added to

[4] http://km.aifb.kit.edu/services/xlike-lexicon/

the cleaned output from the linguistic processing pipelines. This processing was targeted to receive the highest possible recall, so this step provided a noisy output with too many links to Wikipedia articles. This output was then manually verified and cleaned to achieve the clean resource. As before, all the links for detected general concepts were pointing to English Wikipedia articles, but also to other Wikipedias if the respective article existed.

	English	German	Spanish
Automatic NEs	5354	4324	4558
Correct(ed) NEs	3540	3432	3764
Automatic GCs	14522	10347	9719
Correct(ed) GCs	14523	9852	8327

Table 2. Statistics about : NE = Named Entity; GC = General Concept

Just at the first glance the same numbers for Correct(ed) NEs and Correct(ed) GCs would be expected for all three languages since all three texts are parallel and convey the same content. However, the reason for this discrepancy lays in the fact that in different languages, different translation strategies were used (e.g. en: "people of Iran" vs. de: "Iraner", where "Iran" is a NE and "Iraner" is not) and these different solutions lead to different overall counts. The same applies to the wordings or "lexical groundings" in different languages, that doesn't have to refer to the same concepts explicitly in all languages.

```
<sentence id="1">
  <text>An internationally renowned Iranian filmmaker, Mohsen Makhmalbaf, outraged many Iranians by
accepting an invitation to the Jerusalem Film Festival in Israel this month.M</text>
  <tokens>
    <token pos="Z" end="2" lemma="1" id="1.1" start="0">An</token>
    <token pos="RB" end="18" lemma="internationally" id="1.2" start="3">internationally</token>
    <token pos="JJ" end="27" lemma="renowned" id="1.3" start="19">renowned</token>
    <token pos="NP00V00" end="35" lemma="iranian" id="1.4" start="28">Iranian</token>
    <token pos="NN" end="45" lemma="filmmaker" id="1.5" start="36">filmmaker</token>
    <token pos="Fc" end="46" lemma="," id="1.6" start="45">,</token>
    <token pos="NP00SP0" end="64" lemma="mohsen_makhmalbaf" id="1.7" start="47">Mohsen_Makhmalbaf</token>
    <token pos="Fc" end="65" lemma="," id="1.8" start="64">,</token>
    <token pos="VBD" end="74" lemma="outrage" id="1.9" start="66">outraged</token>
    <token pos="PRP" end="79" lemma="many" id="1.10" start="75">many</token>
    <token pos="NP00V00" end="88" lemma="iranians" id="1.11" start="80">Iranians</token>
    <token pos="IN" end="91" lemma="by" id="1.12" start="89">by</token>
    <token pos="VBG" end="101" lemma="accept" id="1.13" start="92">accepting</token>
    <token pos="Z" end="104" lemma="1" id="1.14" start="102">an</token>
    <token pos="NN" end="115" lemma="invitation" id="1.15" start="105">invitation</token>
    <token pos="TO" end="118" lemma="to" id="1.16" start="116">to</token>
    <token pos="DT" end="122" lemma="the" id="1.17" start="119">the</token>
    <token pos="NP00G00" end="146" lemma="jerusalem_film_festival" id="1.18"
start="123">Jerusalem_Film_Festival</token>
    <token pos="IN" end="149" lemma="in" id="1.19" start="147">in</token>
    <token pos="NP00G00" end="156" lemma="israel" id="1.20" start="150">Israel</token>
    <token pos="DT" end="161" lemma="this" id="1.21" start="157">this</token>
    <token pos="NN" end="167" lemma="month" id="1.22" start="162">month</token>
    <token pos="Fp" end="168" lemma="." id="1.23" start="167">.</token>
  </tokens>
</sentence>

...

<node type="entity" class="PERSON" displayName="Mohsen Makhmalbaf" id="E1">
  <mentions>
    <mention sentanceId="1" id="E1.1" words="Mohsen Makhmalbaf">
      <mention_token id="1.6" />
    </mention>
    <mention sentanceId="8" id="E1.2" words="Mohsen Makhmalbaf">
      <mention_token id="8.9" />
    </mention>
    <mention sentanceId="9" id="E1.3" words="Mohsen Makhmalbaf">
      <mention_token id="9.3" />
    </mention>
    <mention sentanceId="16" id="E1.4" words="Mohsen Makhmalbaf">
      <mention_token id="16.24" />
    </mention>
    <mention sentanceId="20" id="E1.5" words="Mohsen Makhmalbaf">
      <mention_token id="20.20" />
    </mention>
  </mentions>
</node>

<node type="entity" class="PRODUCT" displayName="Jerusalem Film Festival" id="E2">
  <mentions>
    <mention sentanceId="1" id="E2.1" words="Jerusalem Film Festival">
      <mention_token id="1.16" />
    </mention>
    <mention sentanceId="3" id="E2.2" words="Jerusalem Film Festival">
      <mention_token id="3.4" />
    </mention>
  </mentions>
</node>

<node type="entity" class="LOCATION" displayName="Israel" id="E3">
  <mentions>
    <mention sentanceId="1" id="E3.1" words="Israel">
      <mention_token id="1.20" />
    </mention>
    <mention sentanceId="5" id="E3.2" words="Israel">
      <mention_token id="5.32" />
    </mention>
    <mention sentanceId="7" id="E3.3" words="Israel">
      <mention_token id="7.18" />
    </mention>
    <mention sentanceId="7" id="E3.4" words="Israel">
      <mention_token id="7.16" />
    </mention>
    <mention sentanceId="9" id="E3.5" words="Israel">
      <mention_token id="9.18" />
    </mention>
    <mention sentanceId="11" id="E3.6" words="Israel">
      <mention_token id="11.10" />
    </mention>
    <mention sentanceId="12" id="E3.7" words="Israel">
      <mention_token id="12.33" />
    </mention>
  </mentions>
</node>

<node type="word" displayName="filmmaker" id="W23">
  <mentions>
    <mention sentenceId="1" id="W23.1" words="filmmaker">
      <mention_token id="1.5" />
    </mention>
  </mentions>
  <descriptions>
    <description URI="10088390-n" displayName="film_maker,filmmaker,film_producer,movie_maker"
knowledgeBase="WordNet-3.0" />
    <description URI="&amp;%Position+" knowledgeBase="SUMO" />
    <description URI="Mx4rvssetJwpEbGdrcN5Y29ycA" knowledgeBase="OpenCYC" />
  </descriptions>
</node>

<node type="word" displayName="month" id="W25">
  <mentions>
    <mention sentenceId="1" id="W25.1" words="month">
      <mention_token id="1.22" />
    </mention>
  </mentions>
  <descriptions>
    <description URI="15209413-n" displayName="calendar_month,month" knowledgeBase="WordNet-3.0" />
    <description URI="&amp;%Month=" knowledgeBase="SUMO" />
    <description URI="Mx4rvVjAKZwpEbGdrcN5Y29ycA" knowledgeBase="OpenCYC" />
  </descriptions>
</node>

...

<annotation displayName="Mohsen Makhmalbaf" entityId="E1" weight="1.000">
  <descriptions>
    <description URL="http://en.wikipedia.org/wiki/Mohsen_Makhmalbaf" lang="en"/>
    <description URL="http://de.wikipedia.org/wiki/Mohsen_Makhmalbaf" lang="de"/>
    <description URL="http://es.wikipedia.org/wiki/Mohsen_Makhmalbaf" lang="es"/>
  </descriptions>
  <mentions>
    <mention sentenceId="1" words="Mohsen Makhmalbaf"/>
    <mention sentenceId="8" words="Mohsen Makhmalbaf"/>
    <mention sentenceId="9" words="Mohsen Makhmalbaf"/>
    <mention sentenceId="16" words="Mohsen Makhmalbaf"/>
    <mention sentenceId="20" words="Mohsen Makhmalbaf"/>
  </mentions>
</annotation>
```

```
<annotation displayName="Jerusalem Film Festival" entityId="E2" weight="1.000">
  <descriptions>
    <description URL="http://en.wikipedia.org/wiki/Jerusalem_Film_Festival" lang="en"/>
    <description URL="http://es.wikipedia.org/wiki/Festival_de_Cine_de_Jerusalén" lang="es"/>
  </descriptions>
  <mentions>
    <mention sentenceId="1" words="Jerusalem Film Festival"/>
    <mention sentenceId="3" words="Jerusalem Film Festival"/>
  </mentions>
</annotation>

<annotation displayName="Israel" entityId="E3" weight="1.000">
  <descriptions>
    <description URL="http://en.wikipedia.org/wiki/Israel" lang="en"/>
    <description URL="http://de.wikipedia.org/wiki/Israel" lang="de"/>
    <description URL="http://es.wikipedia.org/wiki/Israel" lang="es"/>
  </descriptions>
  <mentions>
    <mention sentenceId="1" words="Israel"/>
    <mention sentenceId="5" words="Israel"/>
    <mention sentenceId="7" words="Israel"/>
    <mention sentenceId="9" words="Israel"/>
    <mention sentenceId="11" words="Israel"/>
    <mention sentenceId="12" words="Israel"/>
  </mentions>
</annotation>

<annotation displayName="Film director" entityId="W23" weight="1.000">
  <descriptions>
    <description URL="http://en.wikipedia.org/wiki/Film_director" lang="en"/>
    <description URL="http://de.wikipedia.org/wiki/Filmregisseur" lang="de"/>
    <description URL="http://es.wikipedia.org/wiki/Director_de_cine" lang="es"/>
</descriptions>
  <mentions>
    <mention sentenceId="1" words="filmmaker"/>
  </mentions>
</annotation>

<annotation displayName="Month" entityId="W25" weight="1.000">
  <descriptions>
    <description URL="http://en.wikipedia.org/wiki/Month" lang="en"/>
    <description URL="http://de.wikipedia.org/wiki/Monat" lang="de"/>
    <description URL="http://es.wikipedia.org/wiki/Mes" lang="es"/>
  </descriptions>
  <mentions>
    <mention sentenceId="1" words="month"/>
  </mentions>
</annotation>
```

Figure 2. Example of an English sentence from RECSA annotated for NEs and GCs

Once completed RECSA can be used as the Golden Standard for evaluating cross-lingual semantic annotation and it will be available through META-SHARE[5] language resources sharing platform.

However, since the tasks in the Y3 of the project demand evaluation of further processing steps (e.g., verbal frames in WP2, links to other conceptual spaces in WP3), we are planning to expand the RECSA v1 into RECSA v2.

[5] http://www.meta-share.eu

3 Evaluation of NERC using RECSA

With the availability of the RECSA resource, a standard evaluation methodology for cross-lingual semantic annotation can be conducted. Different semantic annotation systems can use RECSA for measuring their quality since all NEs and all GCs found in the documents are linked to the conceptual space (Wikipedia). The cross-lingual annotation can be evaluated by the number of detected links to the same Wikipedia article by a new system, in comparison to the links existing in RECSA in any of different three languages. This way, the robustness of a method in regard of its performance for different languages can also be measured.

Since the NEs and GCs are annotated inside the text documents by using stand-off XML markup, it is straightforward for evaluators to count all links within the `<node type="entity"…>`, `<node type="word"…>` and `<annotation displayName="…>` elements and automatically compare and count the outputs.

To evaluate the NERC performance with RECSA as new Golden Standard, we used the initial NE processing run through WP2 pipelines for English, German and Spanish. The statistics is presented in Table 3.

	LOC			ORG			PER			MISC			AVERAGE		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
EN	0,727	0,855	0,786	0,291	0,400	0,337	0,546	0,661	0,446	0,198	0,318	0,371	0,440	0,559	0,446
DE	0,836	0,617	0,710	0,236	0,134	0,171	0,301	0,249	0,203	0,034	0,002	0,004	0,352	0,251	0,007
ES	0,764	0,684	0,722	0,453	0,264	0,334	0,479	0,791	0,470	0,185	0,260	0,334	0,470	0,500	0,401

Table 3. Statistics of Named Entity evaluation using RECSA as the Golden Standard

This table gives a clear overview of the quality of the NERC tools used in shallow linguistics processing tested on a RECSA as the Golden Standard instead of CoNLL methodology. Since RECSA v1 was build up with English, German and Spanish texts only, the performance of NERC modules for other XLike languages could not be evaluated. For RECSA v2 we plan the expansion with other languages, we will be able to perform this evaluation for the whole set of XLike languages.

Testing the NERC systems with this changed methodology that uses RECSA as the real life golden standard, the results for all three languages are clearly below what was reported in the First benchmarking report (D7.3.1). What particularly draws attention is a poor performance of the German module. By closer investigation we found an error in output of the German pipeline which originated from a bug introduced in programming the new version of XLike XML schema. This error (dropping out of characters with diacritics) contributed largely to lower recall in German in all categories, but particularly in Miscellaneous category, thus decreasing the overall F1. It didn't affect category Location a lot because most of the location names belong to foreign location names, thus not using German diacritics frequently.

Although the category MISC is heterogeneous and covers NEs that do not fit into Location, Organization or Person, for English and Spanish it performs better than category Organization in this evaluation. One of the reasons for such results could be explained by the fact that English NERC modules were trained for higher recall, leaving the precision at lower grades. This also had an effect to the overall F1 measure. However, we will check the performance of modules for all languages again to investigate what could be the reason for this drop in performance.

The Language identification task was not evaluated separately in this evaluation campaign because during the development cycle of WP2 pipelines and WP3 services in Y2, we were still using the predefined information on language identity from the beginning of the pipeline. This module as well as tokenization module will be evaluated in the extrinsical evaluation campaign that will encompass the whole XLike processing pipelines once they are put together.

Regarding the deep linguistic processing, we performed the evaluation of parsing in D7.3.1, while for semantic role labeling the evaluation will be possible only after the RECSA v2 is build up. RECSA v2 will have verbal frames also marked, so this type of automatic evaluation will be possible as well.

4 Evaluation of the Final Text Annotation Prototype

The purpose of the Final Text Annotation Prototype described in D3.1.2, section 3.2, is to add to the output of WP2 pipelines the annotations with Wikipedia and DBpedia resources. Although we expected to be able to compare the baseline performance of the Early Text Annotation Prototype (described in D3.1.1) with the semantic annotation provided by the Final Text Annotation Prototype developed later (described in D3.1.2), we do not have this possibility due to the introduction of RECSA v1 Golden Standard resource and due to the change in evaluation methodology. This methodology is closer to the real life situation because in RECSA for evaluation we use the parallel corpus of the same genre (news) and not the parallel corpus of a different genre (legal texs, like in D7.3.1). The homogeneity of legal text have contributed to a very high scores in the baseline performance, that we were not sure whether could be repeated in all scenarios and for all XLike languages.

The detailed description of Final Text Annotation Prototype and all the methods used in processing can be found in D3.1.2. This Final Text Annotation Prototype takes into account the output of WP2 pipelines, annotate words and phrases in the text documents and link them to corresponding Wikipedia pages in any language. This protype is capable to link words and phrases to other semantic knowledge representation resource like DBpedia.

In this document we present evaluation of multi-lingual annotation of links from English, German and Spanish texts with the English Wikipedia as described in D3.1.2, section 3.2. The English Wikipedia is taken as a hub knowledge base, as it is by far the largest and best linked Wikipedia.

For evaluation we used a controlled environment provided by the RECSA v1 Golden Standard (see section 2). For each of the documents the Final Text Annotation Prototype linking with Wikipedia were tested for all three languages (en, de, es). This approach automatically inserted links to the English Wikipedia and to German and Spanish (if existed) and these links were then manually evaluated by marking the correctness of the links to English Wikipedia either as **yes**, **no** or **0**. Values **yes** and **no** marked the correct or incorrect link respectively and **0** marked the absence of link to the respective Wikipedia page. In the processing of this evaluation results we took the conservative approach and treated **0** answers as equal to **no**, so the calculated precision is representing only the completely correct links (i.e, only links marked with **yes**). Also, in this evaluation we calculated Precision, Recall and F1 measure for all links cumulatively taken together, so at this stage of evaluation we didn't calculate different values for NEs and general concepts.

	P	R	F1
EN	0.492	0.492	0.492
DE	0.601	0.428	0.500
ES	0.613	0.410	0.491

Table 3. Statistics of the Final Text Annotation Prototype used for inserting links to respective language Wikipedias

The statistics of extracted annotations, i.e. links to Wikipedia, shown in Table 4 demonstrate much less difference in results for the different languages than in the previous evaluation (D7.3.1). However, the performance is below the numbers reported there. We have to point out that these figures can't be compared strictly, because of the change in methodology, i.e. usage of RECSA. Also, in this evaluation NEs are observed together with general concepts and they are expected to represent a category more heterogeneous than NEs. On top of that, the lower performance of NE modules (as reported in section 3) surely contribute to lower overall performance of this service. The precision of the English service is somewhat below expectation and in this case even below German and Spanish. This has to be further investigated in the following development.

5 Evaluation of early machine translation based semantic annotation prototype

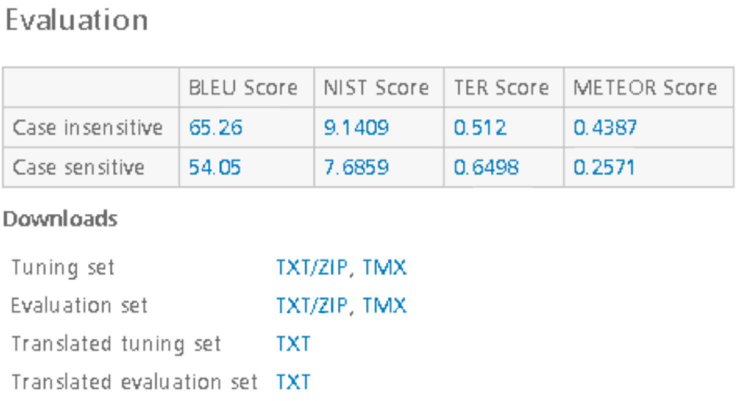
The purpose of the early machine translation based semantic annotation prototype (described more in detail at D3.3.1), is to investigate whether the SMT sytems could be used to translate from natural language into a formal language. This translation would then be used as a semantic annotation of a natural language sentence. Here we present only the summary of evaluation procedure in this task.

The SMT system **En-EnSemRep-Model02** was trained and run on Let'sMT! platform. The translation was also done using the same platform and the result was submitted to evaluation.

In MT community there are two basic types of evaluation of the MT quality: automatic and human.

Automatic evaluation

At the end of the training process the Let'sMT! platform produced automatic evaluation of the trained SMT system using the standard automatic evaluation measures such as BLUE, NIST, TER and METEOR scores.



Evaluation

	BLEU Score	NIST Score	TER Score	METEOR Score
Case insensitive	65.26	9.1409	0.512	0.4387
Case sensitive	54.05	7.6859	0.6498	0.2571

Downloads

Tuning set	TXT/ZIP, TMX
Evaluation set	TXT/ZIP, TMX
Translated tuning set	TXT
Translated evaluation set	TXT

Figure 3. Automatic evaluation of translation quality for En-EnSemRep-Model02 SMT system

The values of these automatic evaluation scores turned out to be good beyond expectations, so we envisaged translations usable also by humans, and not just the machines. However, we still conducted the human evaluation in order to check the quality of the output into FL.

Human evaluation

For the human evaluation in this early prototype of SMT for semantic annotation, we used 1,000 sentences from the test set of 10,000 sentence pairs that was set aside from the training material. This set of 1,000 sentence was translated using **En-EnSemRep-Model02** SMT system and result was submitted to the human evaluation. It was performed by three evaluators, each covering one third of the evaluation set.

The software used for human evaluation was Sisyphos II, an open source MT human evaluation package within the ACCURAT project[6]. We used the Absolute evaluation scenario that uses two categories with several possible values for human judgment: Adequacy and Fluency. Cumulative results of human Absolute evaluation are given in the Table 5.

[6] http://www.accurat-project.eu

Category	Value	Occurences	Percentage
Adequacy	Full content conveyed	209	20.9%
	Major content conveyed	289	28.9%
	Some parts conveyed	270	27.0%
	Incomprehensible	232	23.2%
Fluency	Grammatical	212	21.2%
	Mainly fluent	137	13.7%
	Mainly non fluent	244	24.4%
	Rubble	407	40.7%

Table 4. Results of the human evaluation of translation quality of 1000 English sentences translated into CycL by En-EnSemRep-Model02 SMT system

Interpretation of results from the Table 1 show that human evaluation scored the translation quality of **En-EnSemRep-Model02** SMT system much lower than automatic evaluation. However, numbers show that still a good part of content from English sentences is conveyed into CycL, but it is not done following the strict formal syntax of this FL. This also means that translation from English into CycL, as it is performed by this SMT system, is not immediately applicable for usage where statements with clean and regular CycL syntax are expected.

During Y3 in the continuation of this task we plan also that an **extrinsic evaluation** will be performed, i.e. we will evaluate how the results of this SMT system can be used in further processing steps and how would its usage boost the performance of the whole XLike toolkit.

6 Future evaluation scenarios

Since the tasks in the Y3 of the project demand evaluation of further processing steps (e.g., verbal frames in WP2, links to other conceptual spaces in WP3), we are planning to expand the RECSA v1 into RECSA v2. We see at least three possible directions of expansion by inserting additional manual annotation of:

1) verbal frames,
2) links to Princeton Wordnet v3.0 (incl. SUMO ontology),
3) links to OpenCyc ontology.

In this way the performance of systems that detect (and label) semantic roles or frames, and systems that provide links to other conceptual spaces than Wikipedia, can be evaluated.

We need to be able to evaluate automatical annotations like ones that can be seen in Figures 4 and 5.

```
<frame sentenceId="16" displayName="listen.01" id="F165" tokenId="16.68">
        <argument role="A0:Experiencer" displayName="politician" id="W177"/>
        <argument role="AM-NEG" displayName="not" id="W78"/>
        <argument role="AM-ADV" displayName="even" id="W178"/>
        <descriptions>
                <description URI="02169891-v" displayName="listen" knowledgeBase="WordNet-3.0"/>
        </descriptions>
</frame>
```

Figure 4. Example of frame annotation

```
<node type="word" displayName="art" id="W108">
        <mentions>
                <mention sentenceId="9" id="W108.1" words="art">
                        <mention_token id="9.23"/>
                </mention>
        </mentions>
        <descriptions>
                <description URI="02743547-n" displayName="art,fine_art" knowledgeBase="WordNet-3.0"/>
                <description URI="&amp;%ArtWork=" knowledgeBase="SUMO"/>
                <description URI="Mx4rvVjHuJwpEbGdrcN5Y29ycA" knowledgeBase="OpenCYC"/>
        </descriptions>
</node>
```

Figure 5. Example of links to other conceptual spaces

Also, for RECSA v2 we plan to include all other XLike languages in order to have a Golden Standard capable of measuring the performance of the whole XLike processing toolkit. With RECSA v2 available, we will be able to perform overall automatic extrinsic evaluation of the whole XLike processing platform and clearly notify which modules are contributing and which are subtracting from the overall quality of the processing platform.

7 Conclusion

In this deliverable we have described the evaluation performed on methods and tools developed and described in D2.2.2 and D3.1.2 using the new RECSA Golden Standard. This Resource for Evaluation of Cross-lingual Semantic Annotation was developed during Y2 under task T7.3. Also, evaluation of an Early prototype for semantic annotation using SMT (described in D3.3.1) was presented.

References

[BIZ2006] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia-a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web, 7(3):154–165.

[BO2008] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ACM, pp 1247–1250.

[BUI2008] Paul Buitelaar, Philipp Cimiano. 2008. Ontology learning and population: bridging the gap between text and knowledge, Ios Press, Amsterdam.

[FEL1998] Christianne Fellbaum (ed.) 1998. Wordnet: An electronic lexical database, MIT Press, Cambridge MA.

[HAN2012] Han, A.L.F., Wong, D.F., and Chao, L.S. 2012. LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors. Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Posters, pp. 441–450.

[HOF2013] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and GerhardWeikum. 2013. Yago2: a spatially and temporally enhanced knowledge base from wikipedia. Artificial Intelligence, 194:28–61.

[KRA2012] Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. 2012. Large-scale learning of relation-extraction rules with distant supervision from the web. In The Semantic Web–ISWC 2012, Springer, pp 263–278.

[McN2009] McNamee, P. & Dang, H. T. 2009. Overview of the TAC 2009 Knowledge Base Population Track. In Proceeding of Text Analysis Conference.

[MEL2010] Gerard de Melo, Gerhard Weikum. 2010. Untangling the cross-lingual link structure of Wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics.

[MIH07] Mihalcea, R. and Csomai, A. (2007) Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the 16th ACM Conference on Information and Knowledge management (CIKM'07), Lisbon, Portugal, pp. 233-242.

[MIL08] David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08).

[PAP2002] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL-2002), ACL, pp 311–318.

[RIN2012] D. Rinser, D. Lange, F. Naumann. 2012. Cross-lingual entity matching and infobox alignment in Wikipedia. In Information Systems, Elsevier.

[STE06] Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

[TM02] EF Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In Proceedings of CoNLL-2002.

[TM03] EF Tjong Kim Sang, F De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of CoNLL-2003.