Deliverable D7.3.1

First Benchmarking Report

Editor:	Marko Tadić, UZG
Author(s):	Marko Tadić (UZG), Željko Agić (UZG), Božo Bekavac (UZG), Xavier Carreras (UPC)
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)[1]	Public (PU)
Contractual Delivery Date:	M12
Actual Delivery Date:	M12
Suggested Readers:	All partners of the XLike project consortium and end-users
Version:	1.0
Previous Versions:	0.3, 0.4, 0.5, 0.8
Keywords:	evaluation, linguistic analysis, natural language processing, dependency parsing, semantic role labeling

[1] Please indicate the dissemination level using one of the following codes:
• **PU =** Public • **PP =** Restricted to other programme participants (including the Commission Services) • **RE =** Restricted to a group specified by the consortium (including the Commission Services) • **CO =** Confidential, only for members of the consortium (including the Commission Services) • **Restreint UE =** Classified with the classification level "Restreint UE" according to Commission Decision 2001/844 and amendments • **Confidentiel UE =** Classified with the mention of the classification level "Confidentiel UE" according to Commission Decision 2001/844 and amendments • **Secret UE =** Classified with the mention of the classification level "Secret UE" according to Commission Decision 2001/844 and amendments

Disclaimer

Full Project Title:	Cross-lingual Knowledge Extraction
Short Project Title:	XLike
Number and Title of Work package:	WP7 – Multilingual Linguistic Processing
Document Title:	D7.3.1 – First 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 Y1 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, T2.1.1 and T2.2.1) and early prototype conceptual mapping (WP3, T3.1.1).

Specifically we give results of benchmark tests for (1) methods developed in WP2 for shallow linguistic processing (PoS-tagging, lemmatisation) and deep linguistic processing (dependency parsing) for six XLike languages (en, es, de, ca, sl, zh), and (2) methods developed in WP3 for performance of shallow multi-lingual text annotation tools with a cross-lingual knowledge base, namely Wikipedia for three XLike languages (en, es, de).

This document is the first 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.

Here we present evaluation of the methods described in D2.2.1 and D3.1.1. 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. During year 2 we will analyze the causes for this, and make improvements to meet state-of-the-art accuracies. Evaluation of methods used in early semantic annotation prototype developed within WP3 show performance around the state-of-the-art as reported in references.

Table of Contents

List of Figures

List of Tables

Abbreviations

CoNLL	Conference on Computational Natural Language Learning (http://ifarm.nl/signll/conll)
D	Deliverable
NLP	Natural Language Processing
PoS	Part of Speech tag
XLike	Cross-lingual Knowledge Extraction
WP	Work Package

Definitions

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.

Treebank — A corpus of text documents in which each document is annotated with syntactic and semantic structures. It is used by machine learning methods in NLP in order to train statistical models of language analysis.

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 outcomes are foreseen, but in general it can be said that we expect the performance near the state-of-the-art as reported in referent literature.

In this first benchmarking report we will cover the evaluation of the methods 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 we give results of benchmark tests for:

1) methods developed in WP2 for shallow linguistic processing (PoS-tagging, lemmatisation) and deep linguistic processing (dependency parsing) for six XLike languages (en, es, de, ca, sl, zh);
2) methods developed in WP3 for performance of shallow multi-lingual text annotation tools with a cross-lingual knowledge base, namely Wikipedia for three XLike languages (en, es, de).

This document is the first 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.

2 Evaluation of Linguistic Processing

The goal of WP2 within XLike is to develop methods to analyze documents and extract the entities that appear in the documents, together with their relations. The methods in WP2 should be able to analyze multiple languages – in particular, the six XLike target languages.

The linguistic processing is divided in several layers and at each one of them the evaluation is planned:

1. shallow linguistic processing;
2. deep linguistic processing;
3. extraction of target elements.

The shallow linguistic processing is composed of a pipeline with several steps, i.e. modules, where each adds additional annotation data to the original input text. Each module performs processing of a certain task such as language detection, PoS-tagging, lemmatisation, named entities recognition and classification.

The deep linguistic processing annotates each sentence with the syntactic information in the form of dependency links between words in a sentence, thus providing syntactic hierarchical analysis.

The extraction level of processing annotates the document with the elements from the linguistic structure computed by the previous two layers, such as entitites and relations.

In Y1 only the modules for shallow linguistic processing and parsing as the part of deep linguistic processing were completed up to the stage of early prototypes, so we could evaluate them only.

More detailed report on exact methods used for evaluation of linguistic processing can be found in D2.2.1.

2.1 Evaluation of Shallow Linguistic Processing

The deliverable D2.1.1 documents this set of tools for shallow linguistic processing. In essence, the tools are the following:

1. **Language Identification (id):** Identifies the language of the document and returns the language code.
2. **Tokenization (tok):** Segments and tokenizes the input document. That is, the input is a document free text, and the output is a structure that identifies the sentences and tokens of the document.
3. **Pos Tagging (pos):** Performs lemmatization and part-of-speech disambiguation of sentences, using a statistical tagger.
4. **Named Entity Recognition (ne):** Recognizes the named entities and classifies them according to their semantic type (i.e. Person, Organization, Location, ...), using a statistical entity tagger.

Except the Language Identification method, all other methods are language dependent, meaning that for each language we have a specific language model. This also means that for the evaluation of each module we had to use language dependent procedure following standard CoNLL2009 evaluation scripts in order to stay comparable with state-of-the-art.

During Y1 the evaluation at the shallow linguistic processing for all six XLike languages was performed for tasks:

1. Lemmatisation and PoS-tagging;
2. Named Entity Recognition and Classification.

The tasks of Language identification and Tokenisation were not evaluated for different reasons. For Language identification task an existing system has been used which provides satisfactory results in detecting major languages (en, de, zh, es). Also, it has been previously trained to discriminate between two closely related languages (es and ca) with good results. Regarding Slovenian, in order to discriminate it from other close Slavic languages (e.g. Slovak, Croatian, Serbian…), a system for better discrimination has just been developed [TL12], but it has not been tested yet to the full extent. This will be provided during Y2. At this point the information about the identified language was predefined at the beginning of the pipeline.

The evaluation of tokenisers was not performed because we were facing different tokenization conventions: we couldn't assume that our tools would tokenize the data equally as in each of the CoNLL data sets (note that test CoNLL sets of different languages follow different tokenization conventions). Therefore we decided that in this Y1 evaluation campaign we evaluate tools against test sets where input text is already tokenized.

The accuracy of the **lemmatizers and part of speech (PoS) taggers** was measured using the treebanks described in D2.2.1 as gold standards. The evaluation metric used was accuracy defined as the percentage of lemmas/PoS tags correctly predicted in the test.

Table 1 Evaluation results on the accuracy of Lemmatisation and PoS-tagging

	Lemmas	*Part-of-Speech*
en	96.6	96.6
es	96.3	95.1
de	70.7	88.0
ca	97.1	93.7
zh	100	90.8
sl	97.9	92.5

In some cases, the results are lower than what we expected. This is particularly the case for German, but also in some cases for the other languages. Compared to state-of-the-art for these tasks for some languages (es, ca, sl) we achieved the similar scores (2% difference), but for some we are below reported performance (e.g. German). The reasons for this discrepancy could be found that in some cases, the type of PoS tags of the test set are different than those predicted by the tagger, so we will have to work on the tagset adaptation. Also, during Y2 we will analyze the sources of these lower results and provide solutions.

The **Named Entity Recognition and Classification** evaluation used the methodology of the CoNLL 2003 Shared Task [TM03], but concentrated on four types of entities: locations (LOC), person names (PER), organizations (ORG) and miscellaneous entities (MISC). The evaluation metrics are based on precision and recall at the entity level where they are defined as:

- Precision: the percentage of entities predicted by the system that are correct
- Recall: the percentage of correct entities that are predicted by the system
- F1: the geometric mean between Precision and Recall

For an entity to be considered correctly detected, the words forming the entity and the type of entity have to be correct. Partially recognized entities were considered false.

Table 2 F1-scores of Named Entity Recognition and Classification

	LOC F1	*PER F1*	*ORG F1*	*MISC F1*	*AVG F1*
en	80.5	77.6	61.8	58.0	71.2
es	64.7	80.2	69.6	43.6	69.5
de	61.0	77.4	57.5	51.3	57.8
ca	58.3	71.1	60.1	25.6	58.6
zh	93.5	86.2	73.9	-	88.9
sl	77.1	81.8	-	51.6	72.5

Here we present the summary of the full table reported in D2.2.1, but this one also gives a clear overview of the quality of the NERC tools used in shallow linguistics processing. For some languages some NE categories (for zh MISC and for Sl ORG) were not processed since the gold standard was not annotated using these classes. In Y2 we will fill this gap with providing the better annotated gold standards.

For most of the languages the results are clearly below the best performing systems for respective languages where XLike English and German NERC systems, as they are trained now, would end up in one of the last three positions as reported in the Table 1 of [TM02] and Table 5 of [TM03]. We will have to analyse the source of this discrepancy. One of the first directions is checking the category MISC since the averages for English, German, Spanish and Catalan calculated without this category taken into account, are better from 2 to 7 % points. Most probably this category is introducing a lot of noise in this statistically based techniques and we would like to experiment with the alternative approaches.

However, for these initial versions of shallow linguistics processing pipelines, we can be satisfied that for all six XLike languages we have a common processing infrastructure set in the place and available for processing documents.

2.2 Evaluation of Deep Linguistic Processing

The D2.2.1 reports on the tools developed for this level of processing. In essence these methods perform:

1. **Dependency Parsing (syn):** Performs syntactic disambiguation, producing a dependency tree for each sentence.
2. **Semantic Role Labeling (srl):** Analyzes the predicate-argument structures of the sentence. Specifically, it identifies the lexical predicates of the sentence, and annotates their arguments, each tagged with a semantic role. This task lies in between syntactic and semantic disambiguation.

Following the methodology developed in [NHK+07], we evaluated dependency parsers with the following accuracy measures:

- Unlabeled Attachment Score (UAS): percentage of words with the correct head, ignoring the dependency label;
- Labeled Attachment Score (LAS): percentage of words with correct head and dependency label

For dependency parsing evaluation we use the standard test sets on the treebanks described in D2.2.1. The results are the following:

Table 3 Accuracy of Dependency Parsers using predicted PoS Tags

	UAS	*LAS*
en	86.1	83.0
es	86.6	82.5
de	80.0	75.2
ca	85.8	81.5
zh	80.0	72.0

Since the dependency parsing is highly dependant on the quality of PoS tags, our results are below expected because in some cases the tagsets were not identical to the tagsets expected by the parsers. This contributed to lower results than predicted. In order to check this we did an additional evaluation where we run the dependency parser for English using the correct and by parser expected PoS tags, and we obtained UAS=90.5 and LAS=89.2, a result much more comparable to the state-of-the-art [MCP05,MP06,Car07,NHK07+].

3 Evaluation of Early Text Annotation Prototype

The purpose of the early text annotation prototype described in D3.1.1 is to investigate the performance of shallow linguistic processing annotation tools with a cross-lingual knowledge base, namely Wikipedia. The resulting baseline performance will be compared to the semantic annotation tool developed later in D3.1.2. While this prototype does only annotate word phrases in the text documents and link them to Wikipedia pages in any language, the final annotation prototype will extract subject-predicate-object triples (output of D2.2.1 and D2.2.2) and link them to a semantic knowledge representation like Wikidata or Cyc.

In this document we present evaluation of three approaches for multi-lingual annotation of links with the English Wikipedia as described in D3.1.1:

1. Named Entity Recognition (NER): This approach is based on the Named Entities detected by NERC tools described in D2.1.1. On top of that a simple approach for finding the corresponding Wikipedia pages in the target language (i.e., English) is deployed.
2. Wikipedia Miner Wikifier (WIFI): This approach is trained on existing links in Wikipedia articles to detect similar phrases and links in any text document of the same language as the Wikipedia used for training. Again, a simple approach for finding the corresponding Wikipedia pages in the target language is deployed.
3. Cross-lingual Explicit Semantic Analysis or Wiki Topics (WT): While the first two approach detect word phrases this service links articles by topic to corresponding Wikipedia pages in the target language.

While all implemented tools support many languages for practical reasons and as the proof of a concept, the evaluation was focused on annotating German and Spanish documents only and linking them to the respective English Wikipedia articles. The English Wikipedia is taken as a hub knowledge base, as it is by far the largest and best linked Wikipedia.

The detailed description of Early Text Annotation Prototype and all the methods used in processing and evaluation can be found in D3.1.1. Here we present only the summary of evaluation results.

For evaluation we used a controlled environment in the form of parallel JRC-Acquis corpus [STE2006] with English, German and Spanish 88 documents running in parallel. For each of the documents all three approaches (NER, WIFI, WT) were tested for all three languages (en, de, es).

Each approach automatically inserted links to the English Wikipedia 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 link to Wikipedia disambiguation 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**).

3.1 NER evaluation

The precision of extracted annotations shown in Figure 1 demonstrate difference in results for the different languages. The similarity of the German NER service to the English benchmark service seems to be much lower than to the Spanish one. Even the precision of the English service is somewhat below expectation. This has to be further investigated in the following development.

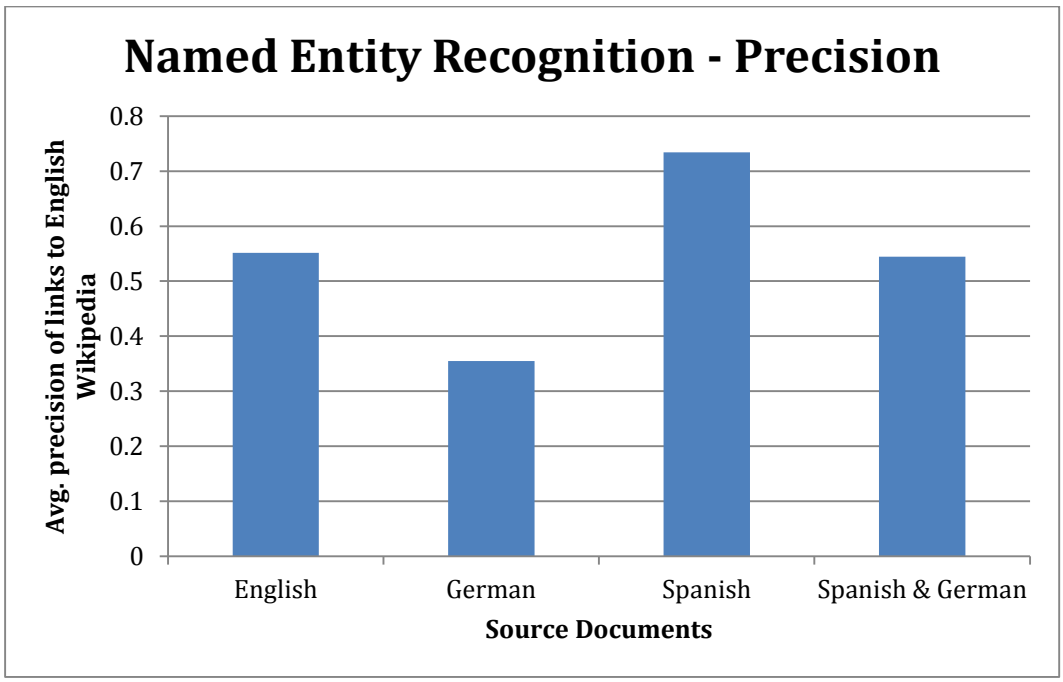

Figure 1. Average precision of NE links to English Wikipedia

3.2 Wikifier evaluation

The German service detects more identical links found by the English service compared to the Spanish service, but for all three languages precision is above 92% which can be considered a very good result for the automatic mapping to the English Wikipedia. This precision will be checked for other XLike languages as well. Considering that a lot more links are extracted compared to the NER services this approach should also be significantly better in terms of recall. This results are in accordance with the state-of-the-art as reported in references [MIH07, MIL08].

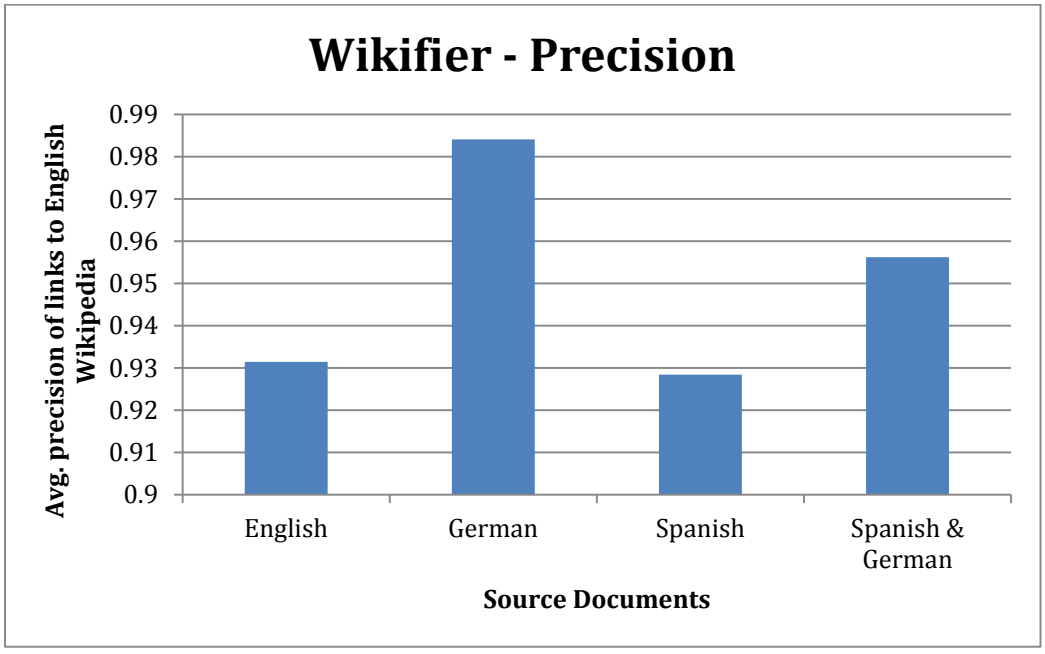

Figure 2. Average precision of Wikifier links to English Wikipedia

3.3 Wiki Topics evaluation

In this evaluation we checked if the top Wikipedia topics (Wiki Topics, WT) associated to the English document is the same as the one associated to the source language (German and/or Spanish) and linked to the English language (for more detailed description of this approach see D3.1.1). Here the precision is lower than in the previous approach and the German system is clearly performing the lowest.

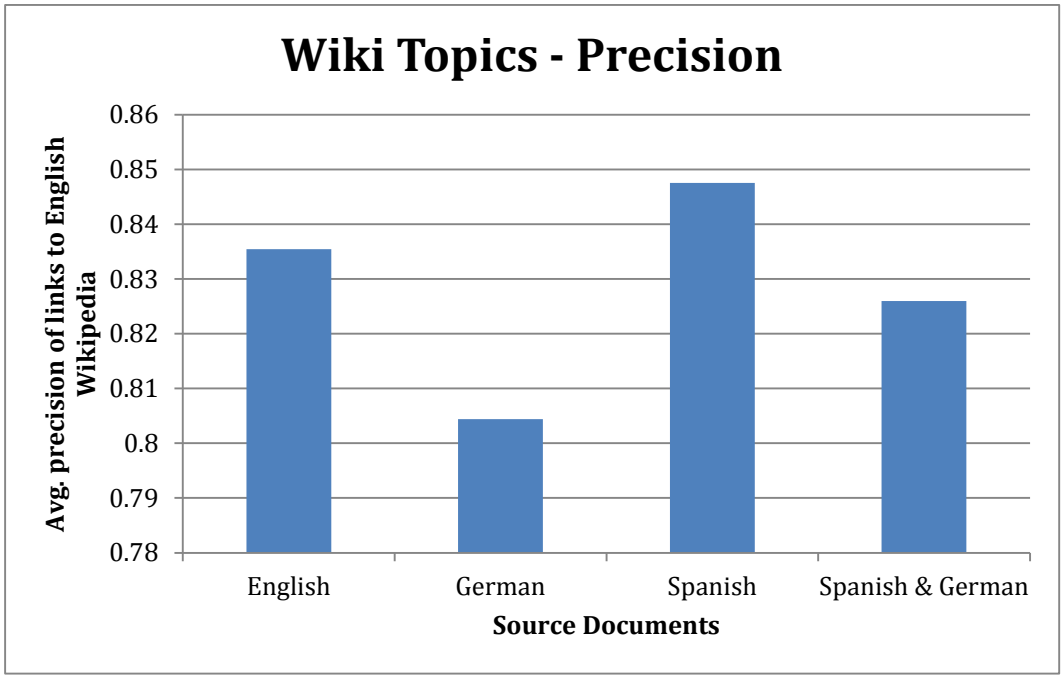

Figure 3. Average precision of CLESA links to English Wikipedia

3.4 Summary of Evaluation

At the end of evaluation we also calculated cumulative precision levels of possible combinations of these three approaches to cross-lingual linking to English Wikipedia articles. The rationale for this is to check whether the any improvement in precision could be achieved by combining two approaches and whether this combining is language dependent or independent. The averaged precisions over the different combination of methods is shown in Figure 4. As it can be seen there, the combination of Wikifier (WIKI) and Wiki Topics (WT) gives the combined precision around 90% and, if proved for other XLike languages, it can be a good ground for further research.

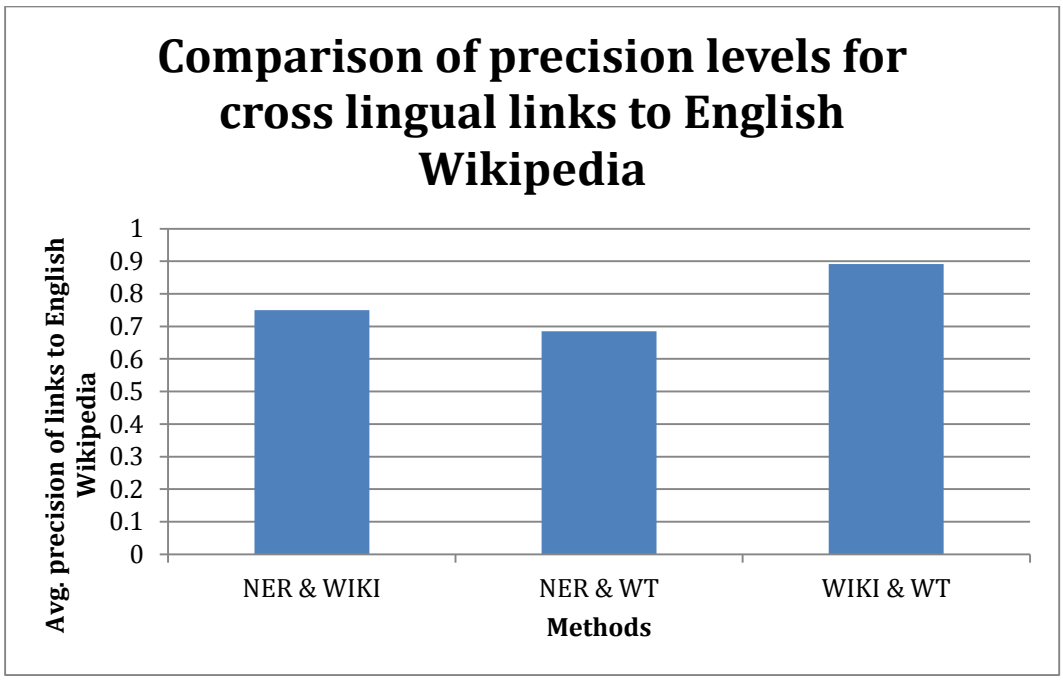

Figure 4. Average combined precisions of all possible combinations of three approaches to cross-lingual linking to English Wikipedia articles

4 Benchmarking

Comparing the evaluation results with the expected benchmarking procedure as described in B1.1.3 of DoW we can say the we have met the following objectives:

1. Objective 1:
    a. Linguistic part of the pipeline (Shallow and Deep Linguistic Processing)
        i. Precision and recall of PoS-tagging, lemmatisation, NERC and parsing
    b. Semantic part of the pipeline (Early Text Annotation Prototype)
        i. Precision of Named entities in relevant knowledge bases (Wikipedia)
    c. Cross-lingual part of the pipeline
        i. Accuracy in cross-lingual linking of documents (Early Text Annotation Prototype)
2. Objective 4:
    a. General aspects
        i. International papers published
            1. Ljubešić-Tiedemann: “Efficient Discrimination Between Closely Related Languages”, COLING2012 oral paper
        ii. Number of languages included in the Shallow Linguistic Processing layer
            1. pipelines developed for all six XLike languages

After Y1, where the Shallow Linguistic Processing pipeline and Early Text Semantic Annotation prototypes were planned, we believe we have met most of the planned objectives.

5 Conclusions

In this deliverable we have described the evaluation performed on methods and tools developed and described in D2.1.1, D2.2.1 and D3.1.1. Also, the benchmarking check with the planned project progress was performed for Objectives 1 and 4.

References

[Car07] Xavier Carreras. Experiments with a Higher-Order Projective Dependency Parser. In Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL, pages 957–961. Association for Computational Linguistics, 2007.

[MCP05] Ryan McDonald, Koby Crammer, and Fernando Pereira. Online Large-Margin Training of Dependency Parsers. In Proceedings of the 43rd ACL, pages 91–98. Association for Computational Linguistics, 2005.

[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).

[MP06] Ryan McDonald and Fernando Pereira. Online Learning of Approximate Dependency Parsing Algorithms. In Proceedings of the 11th EACL, pages 81–88. Association for Computational Linguistics, 2006.

[NHK+07] Joakim Nivre, Johan Hall, Sandra Kubler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The CoNLL 2007 Shared Task on Dependency Parsing. In Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL, pages 915–932. Association for Computational Linguistics, 2007.

[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.

[TL12] J. Tiedemann, N. Ljubešić. Efficient Discrimination Between Closely Related Languages. In Proceedings of COLING2012.

[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.

