SEVENTH FRAMEWORK PROGRAMME
THEME 3
Information and Communication Technologies

PANACEA Project

Grant Agreement no.: 248064

Platform for **A**utomatic, **N**ormalized **A**nnotation and
Cost-**E**ffective **A**cquisition
of Language Resources for Human Language Technologies

D8.2: Tool-based Evaluation of the PANACEA Production Chain

Dissemination Level:	Public
Delivery Date:	December 31, 2012
Status – Version:	V1.1 final
Author(s) and Affiliation:	Vera Aleksić, Chr. Schwarz, Gregor Thurmair (LG), Prokopis Prokopidis, Vassilis Papavassiliou (ILSP), Antonio Toral (DCU)

Relevant PANACEA Deliverables

D3.4	Third version of the Platform
D4.1	Technologies and tools for corpus creation, normalization and annotation
D4.5	Final prototype of the Corpus Acquisition and Annotation Subsystem
D5.2	Aligners Integrated into the Platform
D5.4	Bilingual Dictionary Extractor
D7.4	Third Evaluation Report

Table of Contents

1 Introduction

After the end of the third evaluation cycle, the PANACEA project launched a final work package, dealing with the evaluation of the PANACEA results in industrial contexts. One focus was the usability of the tools in larger practical setups. A given domain (health & safety) and a given language direction (Italian to German) were investigated, and all tools involved from crawling to bilingual lexicon generation were applied.

The main evaluation criterion was usability in industrial contexts; this comprises robustness, effort, and achieved quality of the single tools in comparison with other (state-of-the-art) tools; detailed evaluation was already done in the development cycles, and should not be repeated here.

2 Task Description

The task of ‚tool-based evaluation‘ of the PANACEA tools will look at the single tools to be used in a workflow, in our case domain-adaptation of Machine Translation systems, and look at their quality, and workflow integration. The overall goal in the package ‘Industrial Evaluation’ is to find out how close the components of the PANACEA toolbox are to usability in real situations of industrial production.

It was agreed to use:

- as domain: Health&Safety / Arbeitsschutz / sicurezza sul lavoro, with a focus on the subdomain of construction industry
- as languages: Italian to German

The processing chain to be evaluated, and the evaluation steps, is shown in Fig. 2-1.

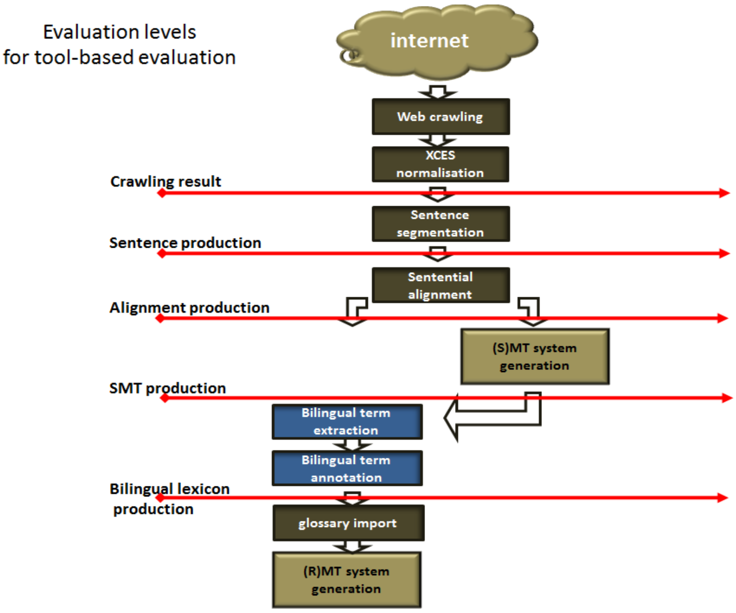


Fig. 2-1: Evaluation levels

Tool-based evaluation will have a closer look on the output of the following components:

- Bilingual crawler. The **crawling result** determines all following steps
- Sentence segmentation and alignment. A list of aligned sentences is what remains from the text preparation results.
- Machine Translation system. We want to assess the quality that can be expected from SMT production.
- Lexicon Extraction. We want to know how useful the bilingual lexicon resulting from this activity really is.

3 Crawling

For crawling, the ILSP Focused Bilingual Crawler was used. Due to temporal restrictions when running the crawler as a web service, it was decided that ILSP would run the crawler themselves.

3.1 Crawler preparation

The crawler needs as input: a list of seed URLs, and a list of seed terms for the topic identifier.

For seed URL collection, to crawl documents in this domain, URLs have been collected pointing to both parallel and monolingual sites. The source of the parallel URLs is mainly:

- European / International (EURLex, OSHA, ILO)
- Swiss (ECAS, SGAS)
- Italian / Regione di Bolzano

Some monolingual sites are added as well, based on the dmoz directory.

For seed term creation, we mainly used a glossary provided by the province of Bolzano, enriched by Eurovoc terms and by terms extracted from en-it and en-de lexicons available at Linguatec.

Both seed URLs and seed terms are given in the Annex.

3.2 Crawler results

For the acquisition process we used a revised version[1] of the Focused Bilingual Crawler developed in WP4. To guide the FBC, we make use of the following resources provided by Linguatec: i) a bilingual topic definition, which was the union of the monolingual topic definitions and ii) the seed URLs of bilingual websites (see the data in the Annex).

After crawling the seed multilingual web sites, the web documents that were relevant to the domain (Health&Safety / Arbeitsschutz / sicurezza sul lavoro) and in the targeted languages (DE, IT) were stored at the nlp.ilsp.gr server. Then, for each selected document, the downloaded HTML file was extracted and the corresponding CesDoc file was created. The next step involved the detection and removal of duplicates. After this stage, the pair detection module examined a) the links to images included in the HTML source and b) the structure of the files to identify pairs of parallel documents.

For each detected pair, a CesAlign file was created, with links pointing to the corresponding CesDoc files. The final output of the FBC is a list of links pointing to the CesAlign files.

The following Table presents the web sites from which the 807 pairs of documents were acquired and the output list for each crawl job.

Website	# of pairs	Output
http://eur-lex.europa.eu	60	http://nlp.ilsp.gr/soaplab2-results/output_eur-lex_list.txt
http://osha.europa.eu/	254	http://nlp.ilsp.gr/soaplab2-results/output_osha-europa_list.txt
http://wegleitung.ekas.ch and http://guida.cfsl.ch	1	http://nlp.ilsp.gr/soaplab2-results/output_wegleguida_list.txt
http://www.assoimprenditori.bz.it	4	http://nlp.ilsp.gr/soaplab2-results/output_assoimprend_list.txt

[1] The main enhancements concern i) the introduction of an additional method for pair detection based on the links and filenames of the images included in the HTML source and ii) the required modification of the FBC in order to visit two websites (e.g. http://www.provincia.bz.it and http://www.provinz.bz.it) instead of "staying" in and crawling only one multilingual website

http://www.ekas.admin.ch	13	http://nlp.ilsp.gr/soaplab2-results/output_ekas-admin_list.txt
http://www.entsendung.admin.ch	8	http://nlp.ilsp.gr/soaplab2-results/output_entsendung-admin_list.txt
http://www.ilo.org	2	http://nlp.ilsp.gr/soaplab2-results/output_ilo_list.txt
http://www.provincia.bz.it and http://www.provinz.bz.it	232	http://nlp.ilsp.gr/soaplab2-results/output_provin_list.txt
http://www.sicuro.ch	1	http://nlp.ilsp.gr/soaplab2-results/output_sicuro_list.txt
http://www.ssst.ch/	3	http://nlp.ilsp.gr/soaplab2-results/output_ssst_list.txt
http://www.suva.ch	229	http://nlp.ilsp.gr/soaplab2-results/output_suva_list.txt

Table 3-1: Crawling Result

Overall, the crawler delivered about 1600 documents of the health & safety domain, about 800 document pairs, containing 1.40 million tokens in Italian, and 1.21 million tokens in German.

3.3 Crawler performance Evaluation

For easier processing, the crawler output file structures were adapted, and unique single filenames were created.

3.3.1 Evaluation questions

The crawler used in the evaluation has four functions; each of them was inspected.

1. Parallelism: It looks for parallel documents in the web. While it is hard to talk about recall (esp. how many parallel documents were missed), however, it can be evaluated if the resulting documents are parallel or not. So the question was: How many documents were parallel / somewhat parallel (i.e. more than 20% of the text is parallel) / not parallel?

2. Topic Identification: As the crawler is a focused crawler, it is relevant how accurate the topic detection of the crawler is. The selected topic was Health&Safety, with a focus of H&S in the construction industry. The question was: How many documents fit to the domain / fit at least partially / do not fit = have a different topic?

3. Language Identification: The crawler determines the language of a document, and marks out-of-language paragraphs. The question was how many paragraphs (textual, not boilerplate) have correct language identification (correctly/incorrectly marked as ‚ool‘, or correctly/incorrectly not marked as ‚ool‘).

4. Boilerplate identification: The crawler marks boilerplates, i.e. document parts which do not belong to the text flow. The question is: How many ‚good‘ texts disappear in boilerplates, and how many ‚bad‘ texts stay in the text flow?

3.3.2 Evaluation Data

Of the 800 document pairs, 100 were randomly selected for manual inspection, coming from all crawled directories. One evaluator was given the task, so there may be slight modifications in the numbers as there are always unclear cases.

Evaluation was supported by a modification of the XCES style sheet, provided by ILSP, so the different kinds of paragraphs (boilerplates, ooi, ool etc.) were marked in different colours; this made the evaluation much simpler.

3.3.3 Evaluation Results

1. Evaluation of parallelism

As mentioned, it could not be evaluated how many parallel documents were *not* found by the crawler. For the ones found, results are given in Table 3-2.

no documents evaluated	103	
parallel	94	91,2%
parallel parts	5	4,8%
nonparallel	4	3,8%

Table 3-2: Parallelism

This shows that the crawler returns sufficiently good material for further processing (about 95% of the crawled data are usable for building parallel resources).

Of course the results depend on the quality of the seed URLs provided by the system user; finding good parallel URLs is not a completely simple task.

2. Evaluation of domain specificity

Next, the topic information returned by the crawler was inspected, both on the Italian and the German side. The result is given in Table 3-3.

language	it		de	
no documents evaluated	103		103	
domain-specific (fully or partially)	79	76,6%	80	77.6%
irrelevant	24	23,3%	23	22,3%

Table 3-3: Domain specificity

Reported results of topic identification in industrial contexts sometimes score higher; however they neglect a main factor of influence, namely the distance between training data and recognition data. In ‘real world‘, scores between 75% and 85% are realistic to assume, with the current results scoring at 77% for at least partially relevant documents.

It should be mentioned that the quality of topic identification strongly depends on the quality of the seed terms; by inspecting the results, modifying the seed terms, and re-crawling results still could be improved.

3. Evaluation of language identification

As errors, only misinterpretation in ‚good‘ paragraphs (no crawlinfo attribute) were counted:

- language different from the one taken as default
- language claimed to be different but is not (or is recognised wrongly)

Results are given in Table 3-4.

language	it	de
no errors in Lang. Identification	13	65
total no ‚good‘ paragraphs	5210	4749

percentage	99,75%	98,63%

Table 3-4: Language Identification

80% of the errors were found in one single document (eurlex-212 (de) / eurlex-213 (it)); without it, figures would be even better (99,67% for de and 99,95% for it). The eurlex-212/213 document contains a lot of chemical substances which are difficult to assign a single language to. The rest of errors is mainly due to paragraphs containing several languages within themselves.

4. Errors of boilerplate recognition

Boilerplate treatment was evaluated as well. The main difference was made between ‚good‘ and ‚bad‘ paragraphs, i.e. paragraphs with and without a ‚crawlinfo‘ attribute. Errors were counted if

- a ‚good‘ text part was marked as boilerplate, or if
- a boilerplate content was found in the ‚good‘ text, i.e. was not recognised.

Also, cases were evaluated where a boilerplate in one language was considered a non-boilerplate in the other language (which will create problems in alignment).

Results are given in the Table 3-5. Again, 100 documents per language were inspected.

language	it	de
total number of paragraphs	23178	23176
wrong boilerplates	2326	2591
percentage of correct recognition	89,96%	88,82%

Table 3-5: ‘Boilerplate’ removal

Basically the error rate here is about 10%. It should be noted that there are different strategies for boilerplate removal which can be followed in treating the texts:

- One option is to remove everything which does *not* belong to the *text* (HTML information around the real text); this is the ‘classical‘ boilerplate approach.
- Another option is to remove everything which is irrelevant for MT sentence alignment; this goes beyond the first approach as it also removes short text chunks, copyright issues and other paragraphs.

The errors in this section are mainly due to this difference; i.e. if the text itself contains paragraphs which are not usable for MT alignment.

It should be noted that the crawler has different ‘crawlinfo’ values to cope with this problem: The ‘boilerplate’ value is set if a classical boilerplate chunk is identified[2]. The other values, ‘out-of-language’, and in particular ‘out-of-length’, give the users the option to integrate the respective texts into the ‘good’ text section or not. In the current tests, they had been excluded, which may not be the best option if the complete text flow is to be kept (e.g. for indexing or information extraction purposes).

In particular the out-of-length attribute has a critical influence here: It causes text parts like headings, or some enumeration / list elements to disappear in the resulting text. It may be worthwhile to reconsider this parameter, esp. if the crawler is to be used for monolingual applications in the information extraction domain.

[2] following Kohlschütter et al. 2010

3.3.4 Recommendations

Overall, the impression is that the error rates of the different components are low enough to enable the use of the PANACEA Focused Bilingual Crawler in industrial setups; the crawler performance matches industrial usability requirements.

In future versions, improvements of the following kind could be imagined:

- Names like ‘ekas-admin_f\ae9bfeef-bcac-44d3-a356-c0b4d20e2962\xml\86.html‘ may have a more user-friendly correspondence. Also, in the alignment files, it would help if the order could reflect the source / target language assignment of a document pair.
- Strategy of the topic identifier: Beyond co-occurrence of seed terms, more sophisticated models could be used. This will increase crawling / processing time but may reduce the time needed to inspect and revise the crawled data.
- Boilerplate: The crawlinfo values must be considered with care. A strategy where all ‘crawlinfo’ segments are removed may be good for the detection of parallel segments; if the extraction of the whole text flow is intended then the text portions marked with the ‘out-of-length’ parameter should be kept: In cases where format indicators like <heading> or <list item> are found, the text should be kept as text, not as boilerplate.
- As will be shown below, a major issue is the treatment of enumerations (1. … 2. … 3. or 1.2.1, 1.2.2 and similar ones) in sentence boundary detection. This is a formatting element, and should not come into the text part. A special recogniser would be helpful in such cases, to prevent the enumeration from being interpreted as the first word in the following sentence.

4 Sentence Extraction and Alignment

Once domain-relevant parallel documents are identified, the next step is to extract aligned sentences from them. This task consists of two subtasks: sentence segmentation, and sentence alignment.

4.1 Sentence segmentation

The next step in processing to be looked at is sentence segmentation. In order to assess the quality of some of the PANACEA tools in this area, a standard sentence segmentation module was used, namely the one delivered with the Europarl corpus (www.statmt.org/europarl). It was compared to one of the PANACEA sentence segmentisers, the LT-SSplit segmentiser (cf. D4.5). Both the German and the Italian documents were sentence-segmentised with both tools. The results are given in Table 4-1.

No Segments	it	de
Europarl	44.100	37.300
LT-SSplit	58.900	51.600
common		22.700

Table 4-1: Sentence Segmentation results

It can be seen that the SSplit segmentiser produces significantly more sentences than the Europarl segmentiser; only about half of the sentences are identical. A closer look at the differences was taken as a consequence, by inspecting 1000 sentences in each language. The results of this evaluation are given in Table 4-2.

(1000 sentences)	it	de
Europarl correct	129	120
LT-SSplit correct	353	428
both wrong	518	452

Table 4-2: Sentence Segmentation Evaluation

The evaluation shows that the PANACEA tool is significantly more accurate than the Europarl segmentiser, in German more than in Italian.

The rather high number of incorrect segmentations results mainly from two phenomena:

- Treatment of enumerations as parts of a sentence, and not as a formal element[3]
- Interaction of sentence boundaries and paragraph boundaries. While LT-SSplit treats <p> … </p> markups also as sentence boundary, the Europarl segmentiser does not.

In any case, the PANACEA sentence segmentation is clearly competitive in terms of industrial quality. The question is which effect the significant difference in sentence segmentation has on the alignment, as sentences form the basis of alignment.

[3] This leads to mistakes when the numbers are treated as ordinals, and the constituents are moved in translation: (de) '*2. Löwen sehen wir gern*' -> (en) '*We love to see 2. Lions*'

This is considered to be bad writing in handbooks for technical authors, but still occurs frequently in texts.

4.2 Alignment

For alignment, a standard tool provided in the PANACEA toolbox is Hunalign; it was used for the evaluation. Hunalign produces scores for a given alignment; in PANACEA experiments, it has been proven to be the best strategy to take segments with a score higher than 0.4. If this threshold is used, then results as shown in Table 4-3 are obtained:

No Segments	provided	used	in %
Europarl	37600	19.399	51.4%
LT-SSplit	52.600	28.900	54.9%

Table 4-3: Alignment results

This shows that only about 50% of the texts can really be used for parallel corpora. The results of LT-SSplit are slightly better (by 3%) than the baseline Europarl results. However, even in documents considered as parallel at first, many segments are not usable for parallel training.

To find out how correct the alignment is, 1000 sentence pairs of the resulting corpus have been manually inspected; the results are given in Table 4-4.

(1000 sentences)	correct	in %	wrong
Europarl	817	81,7%	183
LT-SSplit	866	86,6%	134

Table 4-4: Correctness of alignment

The result is that 15 to 20 out of 100 alignments are incorrect, which may negatively influence the creation of SMT resources. Again, LT-SSplit performs slightly better than Europarl.

Hunalign is a standard tool used in SMT production; in PANACEA there was no work on alignment planned; the tools were just integrated into the toolbox. However, the result shows that there is room for improvement.

5 SMT System

The parallel data produced by the aligners were used as input for in-domain training of a Moses system.

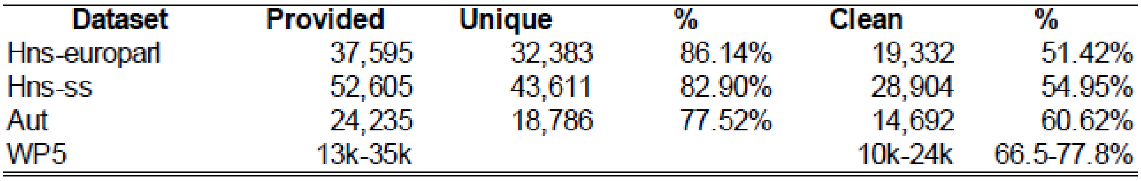
5.1 Preparing Data

The SMT step of the pipeline receives sentence-aligned data (covered in the previous subsection). The data is then tokenised and lowercased using Europarl tools.

The following table give details of the amount of sentences through the preparation process:

- “Provided“ is the amount of sentences output of the aligner without threshold.
- “Unique“ is the amount of sentences after removing duplicate sentence pairs.
- “Clean“ is the amount of sentences after applying the threshold, which removes those sentence alignments with confidence score below 0.4.

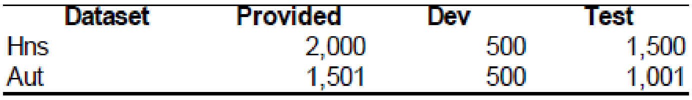
“Hns” is the Health&Safety domain, both with europarl (“Hns-europarl”) and SSplit (“Hns-ss”) sentence segmentation. “Aut” is the automotive domain (used in task 8.3).



Dataset	Provided	Unique	%	Clean	%
Hns-europarl	37,595	32,383	86.14%	19,332	51.42%
Hns-ss	52,605	43,611	82.90%	28,904	54.95%
Aut	24,235	18,786	77.52%	14,692	60.62%
WP5	13k-35k			10k-24k	66.5-77.8%

Table 5-1: Data sets for SMT development

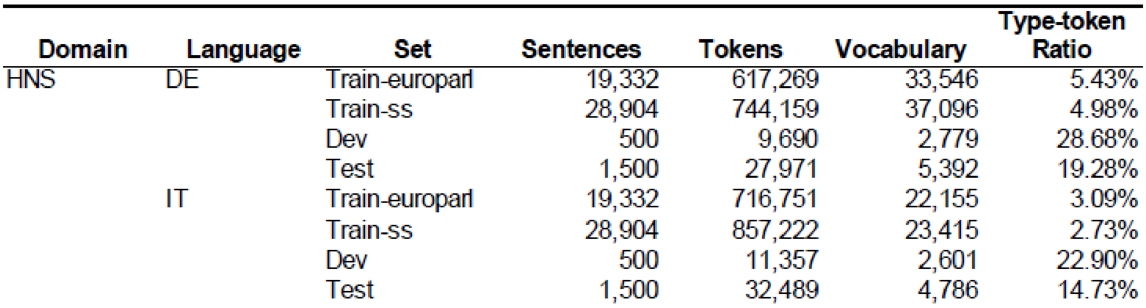
The following table shows the amount of sentences in the development and test datasets. They come from the same dataset and the amount devoted to each set was decided taking into account the findings of Pecina et al., 2012 (more than 500 sentence pairs for development set does not provide a further improvement).



Dataset	Provided	Dev	Test
Hns	2,000	500	1,500
Aut	1,501	500	1,001

Table 5-2: Development and test set for Health&Safety and automotive

Finally, the following table provides quantitative details of each of the datasets. For each of them we show the amount of sentences and tokens as well as the vocabulary size.



Domain	Language	Set	Sentences	Tokens	Vocabulary	Type-token Ratio
HNS	DE	Train-europarl	19,332	617,269	33,546	5.43%
		Train-ss	28,904	744,159	37,096	4.98%
		Dev	500	9,690	2,779	28.68%
		Test	1,500	27,971	5,392	19.28%
	IT	Train-europarl	19,332	716,751	22,155	3.09%
		Train-ss	28,904	857,222	23,415	2.73%
		Dev	500	11,357	2,601	22.90%
		Test	1,500	32,489	4,786	14.73%

Table 5-3: Data Analysis for the Health&Safety domain data

5.2 SMT systems

The MT systems used have been built using Moses (Koehn et al., 2007). For training the systems, training data is tokenized and lowercased using the Europarl tools. The original (non-lowercased) target sides of the parallel data are kept for training the Moses recaser. The lowercased versions of the target sides are used for training an interpolated 5- gram language model with Kneser-Ney discounting using the IRSTLM toolkit (Federico et al. 2011). Translation models are trained on the training corpora (see Section 5.1), lowercased and filtered on sentence level; we kept all sentence pairs having less than 100 words on each side. The maximum length of aligned phrases is set to 7 and the reordering models are generated using parameters: distance, orientation-bidirectional-fe. The model parameters are optimized by Minimum Error Rate Training (Och, 2003) on development sets. For decoding, test sentences are tokenized, lowercased, and translated by the tuned system. Letter casing is then reconstructed by the recaser and extra blank spaces in the tokenized text are removed in order to produce human-readable text.

5.3 Evaluation

The result has been evaluated using automatic metrics as well as human judgement.

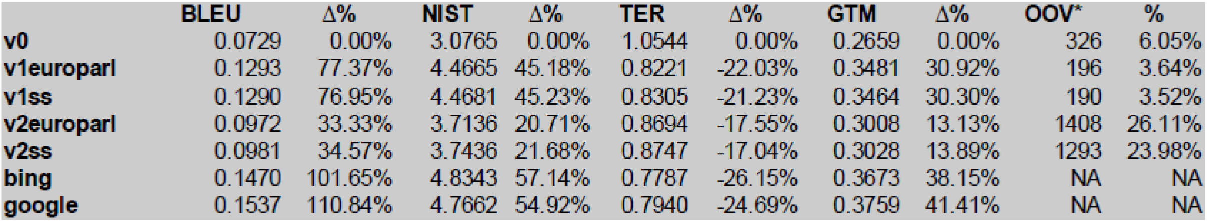
5.3.1 Automatic Evaluation

The systems have been evaluated using a set of state-of-the-art automatic metrics, namely BLEU, NIST, TER and GTM. We report also the size of the vocabulary of the test set and the amount of out of vocabulary (OOV) words.

We have a number of systems according to the data used for training and tuning and the sentence aligner used to split this data:

- The system **v0** is trained and tuned on Europarl, considered to be a general-domain corpus. This is the baseline system to which we will compare our domain-specific systems.
- Systems **v1** are trained on the union of Europarl and the domain-specific data and tuned on domain-specific data.
- Finally, systems **v2** are trained and tuned on domain-specific data only.

Regarding the sentence aligner used, we have two sets of systems. Europarl sentence splitter is used for systems Europarl while SSplit is used for systems Ss. Evaluation results are shown in Tab. 5-4.



	BLEU	Δ%	NIST	Δ%	TER	Δ%	GTM	Δ%	OOV*	%
v0	0.0729	0.00%	3.0765	0.00%	1.0544	0.00%	0.2659	0.00%	326	6.05%
v1europarl	0.1293	77.37%	4.4665	45.18%	0.8221	-22.03%	0.3481	30.92%	196	3.64%
v1ss	0.1290	76.95%	4.4681	45.23%	0.8305	-21.23%	0.3464	30.30%	190	3.52%
v2europarl	0.0972	33.33%	3.7136	20.71%	0.8694	-17.55%	0.3008	13.13%	1408	26.11%
v2ss	0.0981	34.57%	3.7436	21.68%	0.8747	-17.04%	0.3028	13.89%	1293	23.98%
bing	0.1470	101.65%	4.8343	57.14%	0.7787	-26.15%	0.3673	38.15%	NA	NA
google	0.1537	110.84%	4.7662	54.92%	0.7940	-24.69%	0.3759	41.41%	NA	NA

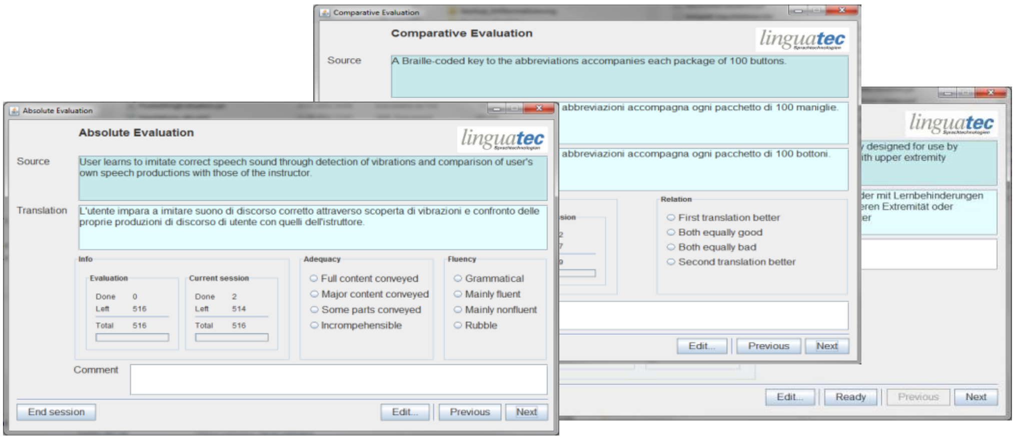
Table 5-4: Automatic Evaluation Results

5.3.2 Human Evaluation

In order to have a second criterion for the quality of the MT output, human evaluation has been performed. 500 Sentences of the test set were inspected:

- A comparative evaluation (between baseline (v0) and adapted (v1ss) system was made.
- An absolute evaluation of the adapted (v1ss) system was made.

For evaluation, the COMP and ABS components of the Sisyphus-II tool were used; screenshots are given in Fig. 5-1.



Absolute Evaluation

Source: User learns to imitate correct speech sound through detection of vibrations and comparison of user's own speech productions with those of the instructor.

Translation: L'utente impara a imitare suono di discorso corretto attraverso scoperta di vibrazioni e confronto delle proprie produzioni di discorso di utente con quelli dell'istruttore.

Evaluation		Current session	
Done	0	Done	2
Left	516	Left	514
Total	516	Total	516

Adequacy: ○ Full content conveyed ○ Major content conveyed ○ Some parts conveyed ○ Incrompehensible

Fluency: ○ Grammatical ○ Mainly fluent ○ Mainly nonfluent ○ Rubble

Comparative Evaluation

Source: A Braille-coded key to the abbreviations accompanies each package of 100 buttons.

abbreviazioni accompagna ogni pacchetto di 100 maniglie.

abbreviazioni accompagna ogni pacchetto di 100 bottoni.

Relation: ○ First translation better ○ Both equally good ○ Both equally bad ○ Second translation better

Fig. 5-1: Sisyphos-II evaluation tools

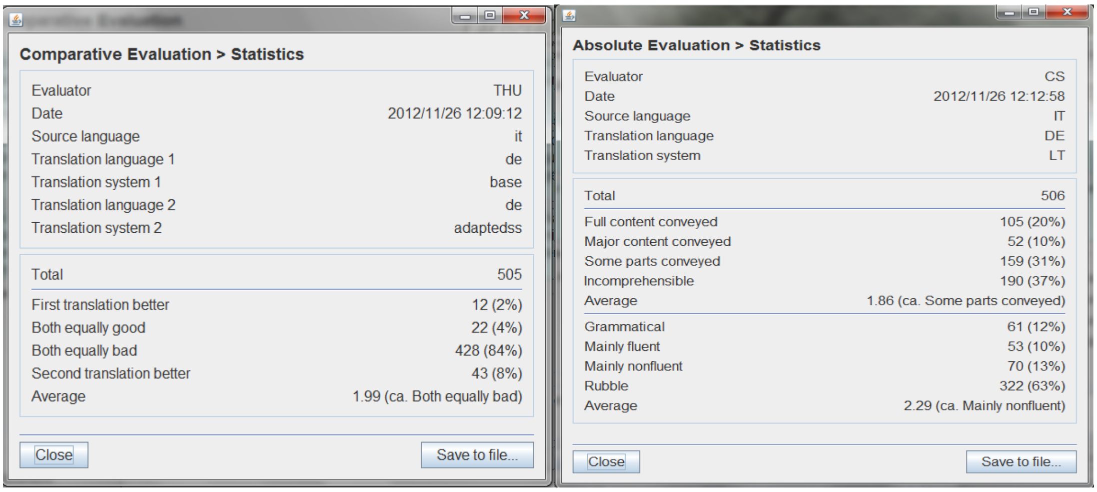
The result of the evaluations is given in Figure 5-2.

Comparative Evaluation showed a 6.13% improvement of the adapted system against the baseline[4]. This is in line with the automatic scores.

Absolute evaluation of the adapted version (v1ss), however, show a rather low MT quality:

- In terms of adequacy (full or major content conveyed), the v1ss system reaches 31.03%
- In terms of fluency (grammatical or mainly fluent), the v1ss system reaches 22.53%

For comparison: In the METAL project, a system was only released when both values were above 70%. In this respect, the MOSES IT-DE system does not meet the release quality.



Comparative Evaluation > Statistics

Evaluator	THU
Date	2012/11/26 12:09:12
Source language	it
Translation language 1	de
Translation system 1	base
Translation language 2	de
Translation system 2	adaptedss
Total	505
First translation better	12 (2%)
Both equally good	22 (4%)
Both equally bad	428 (84%)
Second translation better	43 (8%)
Average	1.99 (ca. Both equally bad)

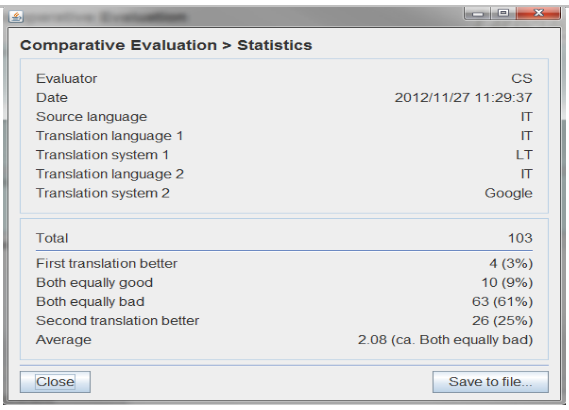
Absolute Evaluation > Statistics

Evaluator	CS
Date	2012/11/26 12:12:58
Source language	IT
Translation language	DE
Translation system	LT
Total	506
Full content conveyed	105 (20%)
Major content conveyed	52 (10%)
Some parts conveyed	159 (31%)
Incomprehensible	190 (37%)
Average	1.86 (ca. Some parts conveyed)
Grammatical	61 (12%)
Mainly fluent	53 (10%)
Mainly nonfluent	70 (13%)
Rubble	322 (63%)
Average	2.29 (ca. Mainly nonfluent)

Fig. 5-2: Human evaluation: COMP v0vs. v1ss, ABS-v1ss

However, as times have changed, a comparison with the state-of-the-art system (Google) was made; Google is slightly better in automatic scores (cf. table 5-4 above). So, 100 sentences of the test were evaluated in a comparison between v1ss and the Google output. The result is shown in Fig. 5-3.

[4] Calculation is: (Improvements – deteriorations) DIV total_sentences



Comparative Evaluation > Statistics

Evaluator	CS
Date	2012/11/27 11:29:37
Source language	IT
Translation language 1	IT
Translation system 1	LT
Translation language 2	IT
Translation system 2	Google
Total	103
First translation better	4 (3%)
Both equally good	10 (9%)
Both equally bad	63 (61%)
Second translation better	26 (25%)
Average	2.08 (ca. Both equally bad)

Fig. 5-3: Human evaluation: COMP V1ss vs. Google

It can be seen that the Google state of the art system improves by 21.35%; however the majority of cases (61%) is still evaluated negative (both equally bad). So the state-of-the-art system it is not significantly better than the PANACEA system. This is also reflected in the automatic scores.

The complete evaluation result is given in Fig. 5-4, comparing v0 and v11ss.

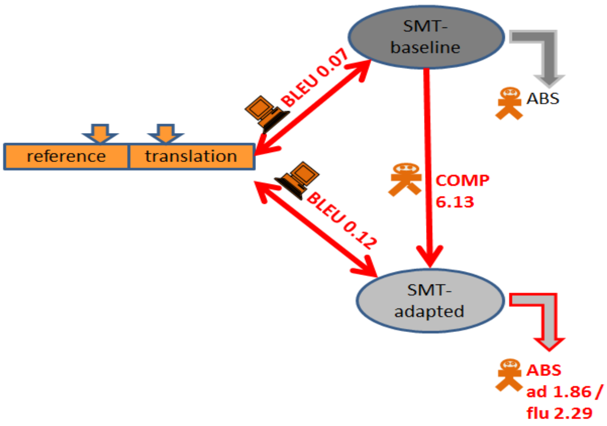


Fig. 5-4: SMT Evaluation results

6 Glossary Production

The last interface to be tested is the output of the glossary production, and the quality of the bilingual lexicon.

The data for evaluation were created with the PANACEA tool LT-P2G, which builds term lists and glossaries from phrase tables.

Two evaluations were made, one for the version v2ss (only in-domain data), and one for the full version (v1ss, containing baseline plus in-domain data).

6.1 Evaluation of v2 data (in-domain only)

As input for this experiment, the phrase table created by the version v2ss was taken; this system contained in-domain data only. The phrase table size is about 1.5 million entries.

To investigate what the best translation probability threshold would be, three runs were made (with P(t|s)=0.6, P(t|s)=0.5, and P(t|s)=0.4). Overall about 3000 entries were extracted from the phrase table data. Table 6-1 shows the result.

Prob.	P(t|s) >0.6	**in %**	0.6>P(t|s)>0.5	**in %**	0.5>P(t|s)>0.4	**in %**
total entries	959		1944		158	
Moses errors	181	18.8%	1029	52.9%	32	20.2%
P2G errors	21	2.2%	15	0.8%	0	0.0%
total errors	202	**21.0%**	1044	53.7%	32	20.2%

Table 6-1: Evaluation of in-domain data

The results show that the restricted quality of the SMT phrase table also influences the glossary extraction component; vice versa the glossary extraction quality depends quite substantially on the quality of the phrase tables. While the P2G error rates are remarkable (less than 2%), an overall error rate of 53.7% for P(t|s)=0.5 is definitely not acceptable in industrial contexts, as every second term candidate would have to be corrected. Even if the threshold is left at P(t|s)=0.6, the error rate is still 21.0%, which is significantly higher than in the development experiments.

Next, it was investigated which effect it has to also look at the reverse translation probability. This is shown in tab. 6-2.

Prob.	P(t|s)>0.6 && P(s|t) > 0.6	**in %**
total entries	536	
Moses errors	69	12.9%
P2G errors	9	1.7%
total errors	78	**14.5%**

Table 6-2: Evaluation of in-domain data: Bidirectional probabilities

It can be seen that the error rate drops by a third, and many incorrect term pairs are filtered out; however the recall is only about half of the recall in the first run. For automatic lexicon production, recall is less important than precision, if no human validation step is included.

6.2 Evaluation of the v1 data (full dataset)

The second evaluation was based on the full data set (version v1ss). The phrase table here contains 104.6 million entries. The run with a basic P(t|s)=0.6 threshold produced a glossary of about 28,300 entries.

Of these, two subsets were manually evaluated:

- a random selection of entries (every tenth entry): 2830 overall
- a selection of only single word entries, 3568 overall

Evaluation results are given in table 6-3.

	every_tenth		single_words		total	
total entries	**2830**	**in %**	**3578**	**in %**	6408	in %
Moses errors	448	15,83	228	6,37	676	10,55
P2G errors	72	2,54	189	5,28	261	4,07
total errors	520	18,37	417	11,65	937	**14,62**

Table 6-3: Evaluation of the full data set

The evaluation shows that the term extraction for the full data set gives better results than for the in-domain data only. After evaluation of about 25% of the result, the error rate comes to 14.6%.

This is still higher than the results during the tests (which had an avg error rate of 9.26%[5]), the main reason being the phrase alignment quality[6]: Higher BLEU scores (as achieved in the data used for component tests, mainly involving English) also produce better glossaries. Lower BLEU scores (in non-English-including directions) reduce the glossary quality.

Although correct term creation in about 85% of the cases still requires manual validation, the overall effort for lexicon production drops substantially, as such a validation is much faster than conventional procedures, as only incorrect proposals need to be deleted to end up with a usable glossary.

6.3 PANACEA It-De-Glossary

From the evaluated data (both v1 and v2), the ‘good’ entries were merged, duplicates were removed, and a POS-annotated, human-validated term list was produced. It contains about 6700 entries.

[5] it should be noted, however, that in tests, Italian-English had already the highest error rate (14.4%), which is close to the figure above.

[6] The high rate of P2G errors in the single word evaluation comes mainly from capitalisation errors in Italian (incorrect lemma of proper names like Assisi etc.)

7 References

- Bel, N., Papavasiliou, V., Prokopidis, P., Toral, A., Arranz, V., 2012: Mining and Exploiting Domain-Specific Corpora in the PANACEA Platform. Proc. 5th BUCC Workshop at LREC 2012, Istanbul
- Federico, M., Bertoldi, N., Cettolo, M., 2011: IRST Language Modeling Toolkit. FBK-irst, Trento
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E.2007: Moses: Open Source Toolkit for Statistical Machine Translation. Proc. ACL, Prague
- Kohlschütter, Chr., et al., 2010: Boilerplate Detection using Shallow Text Features". Proc. WSDM 2010
- Och, F.J., 2003: Minimum Error Rate Training in Statistical Machine Translation. Proc. 41 ACL
- Pecina, P., Toral, A., van Genabith, J., 2012: Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation. Proc. COLING Mumbai
- Thurmair, Gr., Aleksić, V., 2012: Creating Term and Lexicon Entries from Phrase Tables. Proc. EAMT Trento

8 Annex: Crawler Data

8.1 Seed Terms DE

(special files with proper crawler input format have been prepared)

Abbrucharbeiten

Abbruchprogramm

Abriss

Absturzgefahr

Absturzsicherung

Allgemeine Schutzmaßnahme

Anordnung

Anstalt für Arbeitsschutz

Arbeitsaufsicht

Arbeitsauftrag

Arbeitsbedingungen

Arbeitsbedingungen

Arbeitsbühne

Arbeitsgang

Arbeitsmittel

Arbeitsorganisation

Arbeitsplatzgrenzwert

Arbeitsraum

Arbeitsschutz

Arbeitsschutzbestimmung

Arbeitsschutzdienst

Arbeitsseil

Arbeitssicherheit

Arbeitsstätte

Arbeitsstoff

Arbeitsunfall

Arbeitsunfallversicherung

Arbeitsverfahren

Ärztliche Untersuchung

Atemschutzgerät

Auflagefläche

Aufsichtsorgan

Auftragnehmer

Auslösewert

Bauauftrag

Baugewerbe

Baugewerbe

Bauherr

Bauindustrie

Bauindustrie

Bauleiter

Bauphase

Bauprodukt

Bauprojekt

Bauprozess

Baustelle

Baustelle

Baustellenlogistik

Baustellenrichtlinie

Bauunternehmen

Bauunternehmer

Bauwesen

Bemessungsblatt

Berufsgenossenschaft

Berufsunfall

Berufsunfälle

Berufsverband

Biologische Überwachung

Biologischer Arbeitsstoff

Biologischer Arbeitsstofftoleranzwert

Biologischer Grenzwert

Blitzschutzanlage

Brüstung

CE-Kennzeichnung

Chemischer Arbeitsstoff

Dachdecker

Einsatzsicherheitsplan

Elektroanlage

erbgutveränderndernder Arbeitsstoff

Erdungsanlage

Ernste und unmittelbare Gefahr

ESP

Explosionsfähige Atmosphären

Explosionsschutzdokument

Expositionsgrenzwert

Expositionsregister

Fahrbare Leiter

Festigkeitsberechnung

Fluchtweg

Gebäude

Gefahr

Gefährdung

Gefahrenbereich

Gefahrenbeurteilung

gefährlich

Gefahrstoffe

Gerichtsbericht

Gerüst

Gerüstbelag

Gerüstlage

Gesamtstaatliche Beratungskommission für Toxikologie

Gesetzesverletzung im Sachbereich der Arbeitssicherheit

Gesetzlicher Vertreter

Gesundheitsministerium

Gesundheitsschutz am Arbeitsplatz

Gesundheitsüberwachung

Gewerkschaftsbeziehungen

Gleichwertiger Sicherheitsstand

Gleitschutzvorrichtung

Grenzwert

händisches Bewegen von Lasten

Hängeleiter

Hersteller

Höhenarbeiten

Hygienemaßnahme

Impulsförmiger Schall

Individueller Gefahrenschutz

Karzinogen

karzinogener Arbeitsstoff

Kollektiver Gefahrenschutz

Körperschall

krebserregender Arbeitsstoff

Lärm

Lärmbelästigung

Laufsteg

Leiter

Leiter des Arbeitsschutzdienstes

Lieferung und Montage bzw. Einbau

Luftschall

manuelle Handhabung von Lasten

Maschinenrichtlinie

Maximale Arbeitsplatzkonzentration

Mindestinhalt

Mindestvoraussetzungen

Mutagen

mutagener Arbeitsstoff

Notabschaltvorrichtung

Notfall

Oberer Auslösewert

Pausenraum

Periodische Arbeitsschutzsitzung

Periodische Überprüfung

Persönliche Schutzausrüstung

Persönlicher Gehörschutz

Physikalische Einwirkung

PiMUS

Plan für Aufbau, Benutzung und Abbau von Gerüsten

Produktionseinheit

Projektant

PSA

Regeln der Technik

Regelung

Register der Krankheits- und Todesfälle

Rettungsmaßnahmen

Richtlinie

Richtlinie

Risiko

Risikoausschaltung

Risikobeurteilung

Risikobewertung

Risikomappe

Risikominderung

Risikoprävention

Rollgerüst

Rollsteig

Ruhezeit

Schiebeleiter

Schulung

Schutzhelm

Schutzkleidung

Schutzkleidung

Schutzmaßnahme

Schutzmaßnahmen

Seitenschutzgeländer

Selbstsicherndes System

Selbstständige

sichere Instandhaltung

Sicherer Ort

Sicherheits- und Koordinierungsplan

Sicherheits-, Anzeige- oder Kontrollvorrichtungen

Sicherheitsaudit

Sicherheitsausrüstung

Sicherheitsbeauftragte Arbeitnehmervertreter

Sicherheitsfachkraft

Sicherheitsgeschirr

Sicherheitskoordinator in der Ausführungsphase

Sicherheitskoordinator in der Planungsphase

Sicherheitsmerkblatt

Sicherheitsrichtlinien

Sicherheitsstandard

Sicherheitsstufe

Sicherungsseil

SKP

Spezifische Gefahr

Spitzenschalldruck

Ständer

Standfestigkeitsberechnung

Ständiger Beratungsausschuss für Unfallverhütung und Arbeitshygiene

Steckleiter

Strafbestimmungen

Straftat

Strickleiter

Tages-Lärmexpositionspegel

Technische und organisatorische Maßnahmen

Technischer Arbeitsinspektor

Tragbare Leiter

Tragfähigkeit

Transportmittel

Tumorregister

Unfall- und Berufskrankheitenregister

Unfallprävention

Unfallreduzierung

Unfallstatistik

Unfalluntersuchung

Unfallverhütung

Unfallverhütung

Unfallverhütung

Unfallversicherung

Unterer Auslösewert

Verfügung

Verletzung an der Lendenwirbelsäule

Vorankündigung

Vorschriften

Vorsicht

Vorsorgekartei

Wochen-Lärmexpositionspegel

Zeitlich begrenzte Baustelle

Zeitweilige Arbeiten

Zugangsmittel

Zugangsverfahren

Zuständige Betriebsarzt

Zuständige des Arbeitsschutzdienstes

8.2 Seed Terms IT: Sicurezza sul lavoro

Accertamento sanitario

Addetto al servizio di prevenzione e protezione dai rischi

Agente

Agente biologico

Agente cancerogeno

Agente chimico

Agente fisico

Agente mutageno

assicurazione infortuni sul lavoro

Atmosfere esplosive

attività di costruzione

Attrezzatura di lavoro

Calcolo di resistenza

Calcolo di stabilità

Cantiere

cantiere edile

cantiere edile

Cantiere temporaneo

Capacità portante

Cartella sanitaria e di rischio

casco protettivo

cautela

Commissione consultiva permanente per la prevenzione degli infortuni e l'igiene del lavoro

Commissione consultiva tossicologica nazionale

Committente

condizioni di lavoro

Contenuto minimo

Contratto d'opera

Coordinatore della sicurezza per l'esecuzione

Coordinatore della sicurezza per la progettazione

costo degli infortuni

costo di lesioni

costo di malattia

costruttore

Direttiva

Direttiva cantieri

Direttiva macchine

Direttore dei lavori

Dispositivi di protezione individuale

Dispositivi di sicurezza o di segnalazione o di controllo

Dispositivo antiscivolo

Dispositivo di arresto di emergenza

Dispositivo di protezione contro le cadute (dall'alto)

Dispositivo di protezione delle vie respiratorie

Disposizione

disposizione

Documento sulla protezione contro le esplosioni

DPI

edilizia

edilizia dei trasporti

Elemento di appoggio

Eliminazione dei rischi

Emergenza

equipaggiamento di protezione

Esperto della sicurezza

Fabbricante

falegnameria

fase di costruzione

Fase di lavoro

Formazione

Fornitura e posa in opera

Fune di lavoro

Fune di sicurezza

Imbracatura di sostegno

Impalcato

Impianto di protezione contro le scariche atmosferiche

Impianto di terra

Impianto elettrico

Impresa appaltatrice

impresa edile

Inchiesta infortunio

incidente sul lavoro

indumenti di protezione

Indumento protettivo

industria edile

industria edilizia

infortunio sul lavoro

ingegneria civile

Inquinamento acustico

Ispettore tecnico del lavoro

Istituto Superiore Prevenzione e Sicurezza sul Lavoro (ISPESL)

lavoratore del settore edile

Lavoratore/trice autonomo/a

Lavori in quota

Lavori temporanei

lavoro di costruzione

Legale rappresentante

Lesione dorso-lombare

Linee guida

Livello di contenimento

Livello di esposizione giornaliera al rumore

Livello di esposizione settimanale al rumore

Livello di sicurezza equivalente

Locale di lavoro

Locale di riposo

Luogo di lavoro

Luogo sicuro

manutenzione degli edifici

manutenzione sicura

Mappa di rischio

Marcatura CE

Marciapiede mobile

materia di sicurezza

Medico competente

Mezzi di trasporto

Ministero della sanità

Misura generale di tutela

Misura igienica

Misure di emergenza

Misure di prevenzione

Misure di protezione collettiva

Misure di protezione individuale

misure di salvaguardia

Misure tecniche ed organizzative

Monitoraggio biologico

morte bianca

Movimentazione manuale dei carichi

Norma di tutela del lavoro

Norme di buona tecnica

Norme penali

Notifica preliminare

Organo di vigilanza

Parapetto

Passerella

Pericolo

Pericolo grave ed immediato

pericoloso

Periodo di riposo

Piano di calpestio

Piano di montaggio, uso e smontaggio di ponteggi

piano di posa

Piano di sicurezza e di coordinamento

Piano operativo di sicurezza

Piattaforma

PiMUS

ponteggio

ponteggio su ruote

POS

prescrizione

Prescrizione

Pressione acustica di picco

prevenzione

prevenzione degli infortuni

Prevenzione degli infortuni

Procedura di lavoro

processo di costruzione

Produttore

Progettista

progetto di costruzione

progetto edile

Programma di demolizione

Protezione individuale dell’udito

protezione individuale durante il lavoro

PSC

Rapporto giudiziario

Rappresentante dei lavoratori per la sicurezza

Reato

Registro degli infortuni e malattie professionali

Registro dei casi di malattia e di decesso

Registro dei tumori

Registro di esposizione

regolamento

Relazione di calcolo

Relazioni sindacali

Requisiti minimi

Responsabile del servizio di prevenzione e protezione

Riduzione dei rischi

Rischio

Rischio di caduta dall’alto

Rischio specifico

Riunione periodica di prevenzione e protezione dai rischi

RLS

RSPP

Rumore

Rumore impulsivo

Rumore strutturale

Rumore trasmesso per via aerea

sanita del lavoro

Scala a funi

Scala a pioli

Scala a pioli composta da più elementi innestabili

Scala a pioli mobile

Scala a pioli sospesa

Scala a piolo portatile

Scala a sfilo

Scheda dei dati di sicurezza

Servizio di prevenzione e protezione dai rischi

settore edilizio

sicurezza per il settore edile

Sicurezza sul lavoro

Sistema autobloccante

Sistema di accesso

situazione pericolosa

Sorveglianza sanitaria

sostanza pericolosa

statistica sugli incidenti

Superfice di appoggio

Trabattello

Unità produttiva

Uso di attrezzature munite di videoterminali

Valore di azione

Valore inferiore di azione

Valore limite

Valore limite biologico

Valore limite di esposizione

Valore limite di esposizione professionale

Valore superiore di azione

Valutazione del rischio

Verifica periodica

Via di emergenza

Violazione in materia di sicurezza del lavoro

Zona pericolosa

8.3 Seed URLs DE

(first part are parallel URLs)

http://eur-lex.europa.eu

http://de.wikipedia.org/wiki/Arbeitsschutz

http://eur-lex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdoc&lg=de&numdoc=31989L0686&model=guichett

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31989L0656:DE:HTML

http://osha.europa.eu/de

http://www.ekas.admin.ch/index-de.php

http://www.ilo.org

https://experts.tis.bz.it

http://www.entsendung.admin.ch/cms/content/lohn/arbeitssicherheit_de/

http://www.sgas.ch

http://wegleitung.ekas.ch

http://www.issa.int

http://www.sicuro.ch

http://newsletter-vslzh.ch/index.php?id=116&L=2

http://www.ssst.ch/de/

http://www.b-f-a.ch/de/index.asp

http://www.arbeitssicherheitschweiz.ch

http://www.suva.ch/startseite-suva/praevention-suva/arbeit-suva.htm

http://www.provinz.bz.it/arbeit/arbeitsschutz/172.asp

http://www.ingbz.it/content.asp?L=2&IDMEN=182

http://www.assoimprenditori.bz.it/bolzano/notiziario/istituzionale.nsf/codice/576-762?opendocument&lan=de

----

http://www.bgbau.de

http://www.dguv.de/inhalt/index.jsp

http://www.inail.it/Portale/appmanager/portale/desktop?_nfpb=true&_pageLabel=PAGE_HOME_DD

8.4 Seed URLs IT

(first part are parallel)

http://eur-lex.europa.eu

http://it.wikipedia.org/wiki/Sicurezza_sul_lavoro http://eur-lex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdoc&lg=de&numdoc=31989L0686&model=guichett

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31989L0656:IT:HTML

http://osha.europa.eu/it

http://www.ekas.admin.ch/index-it.php

http://www.ilo.org

https://experts.tis.bz.it

http://www.entsendung.admin.ch/cms/content/lohn/arbeitssicherheit_it/

http://www.sgas.ch

http://wegleitung.ekas.ch

http://www.issa.int

http://www.sicuro.ch

http://newsletter-vslzh.ch/index.php?id=116&L=2

http://www.ssst.ch/it/

http://www.b-f-a.ch/it/index.asp

http://www.arbeitssicherheitschweiz.ch

http://www.suva.ch/it/startseite-suva/praevention-suva/arbeit-suva.htm

http://www.provincia.bz.it/lavoro/tutela-del-lavoro/172.asp

http://www.ingbz.it/content.asp?L=1&IdMen=182

http://www.assoimprenditori.bz.it/bolzano/notiziario/istituzionale.nsf/codice/576-762?opendocument&lan=it

----

http://www.lavoro.gov.it/Lavoro/SicurezzaLavoro/

http://www.amsicurezzasullavoro.it/

http://www.ispesl.it/

http://www.intrage.it/rubriche/lavoro/sicurezza/index.shtml

http://www.aifos.eu/

http://it.wikipedia.org/wiki/Testo_unico_sulla_sicurezza_sul_lavoro

8.5 Glossary used:

http://www.provinz.bz.it/arbeit/download/Begriffe_zur_Arbeitssicherheit_12-11-07dt-it.pdf