ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www.accurat-project.eu

Project no. 248347

Deliverable D4.2

Baseline SMT systems enriched with additional data from comparable corpora

Version No. 1.0

29/06/2012

Document Information

Deliverable number:	D4.2
Deliverable title:	Baseline SMT systems enriched with additional data from comparable corpora
Due date of deliverable:	29/02/2012, postponed to 30/06/2012
Actual submission date of deliverable:	29/06/2012
Main Author(s):	Sabine Hunsicker, Yu Chen
Participants:	DFKI
Internal reviewer:	RACAI (Radu Ion, Dan Tufiș)
Workpackage:	WP4
Workpackage title:	Comparable corpora in MT systems
Workpackage leader:	DFKI
Dissemination Level:	**PP**: Restricted to other programme participants (including the Commission Services)
Version:	V1.0
Keywords:	statistical machine translation, language resources, under-resourced languages

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
V0.1	31/05/2012	Draft	DFKI	Initial draft of the deliverable	Delivered for internal review
V0.2	26/06/2012	Review	RACAI	Internal review	Delivered for modifications
V1.0	27/06/2012	Draft	DFKI	Final mofifications	Submitted

EXECUTIVE SUMMARY

Gathering parallel corpora to train machine translation (MT) systems is expensive in terms of money and time. To facilitate the use of MT systems for under-resourced languages, we make use of the comparable corpora and tools developed in the previous project years. We apply the tools to the comparable corpus to extract parallel data and use this extracted data to improve the translation quality of the baseline statistical MT systems reported in D4.1.

Table of Contents

Abbreviations

Table 1 Abbreviations

Abbreviation	Term/definition
MT	Machine Translation
SMT	Statistical Machine Translation
TM	Translation Model
LM	Language Model
BLEU	Bilingual Evaluation Understudy
AC	Acquis Communautaire
EU	European Union
TU	Translation Unit
WMT	Workshop for Statistical Machine Translation
MERT	Minimal Error Rate Training

Introduction

Statistical Machine Translation (SMT) depends on the availability of large quantities of parallel corpora to train their models on. Whereas such corpora are available for the language pairs dominating in current research, such as Arabic-English or German-English, many language pairs suffer from a lack of resources. Since the creation of parallel corpora is expensive, we investigated methods how to extract parallel data automatically from comparable corpora in the ACCURAT project. In this deliverable, we apply the extracted parallel data to our baseline SMT systems to facilitate an improvement in translation quality.
This report is structured as follows: Section 1 explains how we adapt the baseline models with the new data. The data itself is described in detail in Section 2. We report the different experiments we performed in Section 3, and discuss the results in Section 4.

1. Methodology

1.1. Basic Approach

We can use the data from the comparable corpora in manifold ways. When improving SMT systems, we need to look at the two models used in translation: the translation model (TM) and the language model (LM).

The translation model contains the phrase table as well as the reordering model to ensure that the target text generated by the decoder is *equivalent* in meaning to the source text. The language model, however, deals with the *fluency* of the generated text and prefers well-formed translation hypotheses to malformed ones.

The comparable data is used to adapt both models. The resulting new SMT system is evaluated using a pre-defined test set, and translation quality is measured by the automatic metric BLEU. This score is compared to the translation quality of the baseline SMT systems.

In Section 1.2 we explain how we can use the comparable corpus to improve the translation model, whereas in Section 1.3 we give details about adapting the language model.

1.2. Translation Model

In SMT we use parallel data to automatically learn the correspondences between language A and language B by way of translation probabilities. These probabilities form the *translation model*. In phrase-based SMT this model is stored as a phrase table. Each translation option consists of phrase_E in the source language and phrase_F in the target language as well as a number of probabilities, such as the translation probabilities $p(f|e)$ and $p(e|f)$. During decoding all applicable translation options are retrieved from the phrase table. Since there is a large number of translation options, a complete examination of the search space is too computationally expensive, hence, we have to prune our hypotheses, i.e. we drop hypotheses which score worse in comparison to the other hypotheses.

Wrong translations are then caused by two sorts of errors: *model errors* and *search errors*. The latter kind means that there is a better-scoring hypothesis in our search space than the one we ended up with, but we pruned too early or too eagerly. We are not concerned with this kind of error.

Instead we want to fix the model errors. Here we receive incorrect translations because the input sequence contained words unknown to the translation model (*out of vocabulary* errors) or because the probabilities associated with the translation options in questions were not appropriate. We can fix this error by adding more data to our system.

In the following we describe two approaches how we make use of the additional data. First, we can simply extend our already existing training corpus with the new data. Second, in order to increase the influence of the new data, we also use mixture models.

1.2.1. Adding An Additional Corpus

When we have corpus A, the easiest way to retrain our translation model is to simply add corpus B to A in a linear fashion and then rerun the training pipeline. Using this method, there is no further emphasis on the additional data. This usually provides good results especially if we have large out-of-domain or narrow domain parallel corpora and can add general domain data.

1.2.2. Mixture Models

Including additional parallel corpora as training data to a SMT system usually yield an improvement to certain extent. Our hope is that the additional parallel data could bring in new phrases or, more generally, new information that was not contained in the baseline model.

However, the additional texts could also introduce errors that do not exist in the original model. This case is especially more likely to happen when the parallel texts are not translations of each other, for example, when we have misaligned sentences in the comparable corpora. On the other hand, due to various reasons, the added data might not be dominant enough among the other sources of training corpora to help the SMT system to recover from the errors in the base system. Therefore, in addition to a single translation model built from both the parallel corpora and the comparable data as a whole, we also experimented with mixture models that distinguish texts from different sources.

The mixture models start from individual models that are generated separately using the sets of texts from different sources. The most straightforward way is to divide the data into two subsets: the original parallel corpora vs. the aligned texts that were extracted from the comparable corpus. Such a partition may be very close to the baseline model when the sizes of the two subsets differ too much as it would lead to a mixture model that relies on the larger subset. Thus, in order to emphasise and better control the contribution of parallel and comparable data to the final translation, we choose to further divide the original parallel data into separate corpora, from each of which we generate a different translation model. This approach also allows us to understand the influence of each individual corpus, parallel or comparable, in the SMT system and it is especially important when the parallel corpora used in the baseline systems are from very different domains.

Building a standard SMT translation model usually includes the following steps:

1. Data preparation
2. Word alignments
3. Phrase extraction and scoring
4. Translation model construction

As the state-of-the-art word alignment algorithm such as GIZA++ tends to perform poorly for limited amount of data, we generate the word alignments for the mixture model by training over the combination of all the training data, i.e. the parallel data alongside with the extracted sentence pairs from the comparable corpus in order to find sufficient alignment points that are useful for constructing a translation model. Then, after the second step, the word alignments are split into segments corresponding to the individual corpus.

We construct the individual translation models from the word alignments for each corpus following the rest of steps of the standard procedures of phrase-based SMT model training. The models are then sorted by the size of the corresponding training corpora, given the fact that the probabilistic estimation over a larger set of data is usually more reliable.

The other models are appended to the largest model in this sorted order such that only phrase pairs that were never seen before are included. Lastly, we add new features (in the form of additional columns) to the phrase table of the final translation model to indicate each phrase pair's origin. Each new column corresponds to one model, including the original model. If a phrase table entry appears in a model, its feature value in the corresponding column is 2.718, otherwise 1.

Figure 1 shows a few sample entries from the phrase table of a mixture model created in our experiments for English-Latvian translation. The first five columns are the probabilistic

scores estimated in the standard phrase-based SMT training, including the inverse phrase translation probability $\varphi(fee)$, the inverse lexical weighting $lex(f|e)$, the direct phrase translation probability $\varphi(e|f)$, the direct lexical weighting $lex(e|f)$ and the phrase penalty, which is always $e^1 = 2.718$. Following the scheme of defining the phrase penalty, we added three additional columns to the phrase table, corresponding to the three individual models, which have been sorted by size. In this example, the first column refers to the JRC model, the second DGT and the last is for the extracted USFD corpus. The values in these three columns are either 2.718 or 1, indicating whether the phrase pairs exist in the individual models. For example, the last three columns for the phrase pair "*economic approaches*"–"*ekonomiskas metodes*" are 1, 2.718 and 1. This means that this pair is originally from the DGT model and does not appear in the other two.

Table 2 Sample entries from the phrase table of a mixture model for EN-LV.

Source phrase (*e*)	Target phrase (*f*)	Probabilistic scores	Origin markers
economic, political	*ekonomiskās , politiskās*	0.079 0.266 0.011 0.011 2.718	2.718 2.718 2.718
economic, social	*Ekonomiku, sabiedrību*	0.119 0.006 0.008 0.001 2.718	2.718 2.718 2.718
economic, administrative	*ekonomiskiem,administratīviem*	0.238 0.277 0.238 0.001 2.718	2.718 2.718 1
economic, social	*ekonomiskajā , sociālajā*	0.119 0.048 0.001 0.001 2.718	2.718 2.718 1
economic, social ,	*ekonomiskajā , sociālajā ,*	0.079 0.133 0.006 0.003 2.718	2.718 1 2.718
economic activities	*ekonomisko aktivitāšu*	0.205 0.547 0.005 0.001 2.718	2.718 1 2.718
economic downturn	*ekonomikas lejupslīdi*	0.120 0.134 0.017 0.016 2.718	1 2.718 2.718
economic , industrial	*ekonomiskās , rūpnieciskās*	0.120 0.326 0.020 0.006 2.718	1 2.718 2.718
economic , but	*ekonomiskas , bet*	0.119 0.218 0.238 0.008 2.718	2.718 1 1
economic (	*ekonomiskos*	0.238 0.720 0.119 0.010 2.718	2.718 1 1
economic subjects	*ekonomiskajos priekšmetos*	0.406 0.555 0.051 0.001 2.718	1 2.718 1
economic approaches	*ekonomiskas metodes*	0.241 0.004 0.241 0.001 2.718	1 2.718 1
economic relations	*ekonomiskās sadarbības*	0.006 0.003 0.001 0.008 2.718	1 1 2.718
economic threat .	*ekonomisks drauds ,*	0.018 0.024 0.036 0.004 2.718	1 1 2.718

In the mixture model, segments repeated by many sources are considered more probable for translation. On the other hand, unique pieces from some sources may lead us to valuable information, such as terminologies from a particular domain in the comparable corpus. The former case corresponds to phrase pairs with very high probabilities, whereas the latter is still included in the model.

1.3. Language Model

As described in Section 1.1, the language model (LM) makes sure that the selected translation hypothesis is *fluent* in the target language. To achieve this, we use large amounts of monolingual data in the target language to learn a model of probable sequences in the form of *n*-grams. In our decoder we can use multiple language models, so we could train a new language model on the additional data we have acquired and use it in addition to our baseline language models.

This setup is problematic, however, as the data the language models were trained on can differ in domain and style, so that using them in an equal setup will diffuse the importance of the new data, especially if we are adding general domain data to a narrow domain baseline

corpus. To avoid this problem, we are going to interpolate the different language models we trained. The exact procedure is described in the next section.

1.3.1. Interpolating Language Models

To make the best use of the fact that our language models have been trained on different texts, we want to combine them into one and adapt the n-gram probabilities accordingly. Although for example our baseline JRC and DGT language models are out of domain, we do not want to lose the information they contain completely. On the other side, these models are big enough that they can overpower the influence of the new language model that has been trained on much smaller amounts of data. Here we need to adjust the n-gram probabilities so they mirror what we would expect from our target domain.

Combination is done by optimising the perplexity of the interpolated language model on an in-domain development text in the target language. We then receive a lambda for each language model we used; with this parameter we can adjust the probabilities for each n-gram. This way we combine the probabilities from the different language models into one. For details of this approach, please refer to Schwenk & Koehn, 2008[1].

The interpolated language model will then be used for the new SMT system.

[1] Large and Diverse Language Models for Statistical Machine Translation, *Holger Schwenk and Philipp Koehn*, IJCNLP 2008.

2. Data

As seen in Section 1.2, we require parallel data to adapt the translation model. In the following we first describe the data our baseline SMT systems are trained on in Section 2.1, and the comparable data and the ways we used it to enrich the baseline systems in Section 2.2. Beside the training data, we also describe the development and test sets we used in Sections 2.3 and 2.4 respectively.

2.1. Baseline Data

We use the following publicly accessible parallel corpora to set up our baseline SMT systems for the experiments:

- JRC: JRC-Acquis is the parallel corpus collected from the Acquis Communautaire (AC), the total body of European Union (EU) law. It is available for all 22 official EU languages, including at least three of the nine languages that become official in 2004. The texts from JRC-Acquis are sentence aligned automatically.
- DGT: DGT-TM is the multilingual translation memory from the European Commission's Directorate-General for Translation. Being a translation memory, DGT-TM consists of small text segments and their translations, referred to as translation units, TU. The TUs can be sentences or parts of sentences. This memory contains most, although not all, of the documents which make up the Acquis Communautaire, as well as some other documents which are not part of the Acquis.
- SETimes: SETimes is a parallel corpus of news articles in eight Balkan languages and English, originally extracted from the multilingual news website http://www.setimes.com.
- Europarl: The Europarl parallel corpus is extracted from the proceedings of the European Parliament dating back to 1996. It includes versions in 21 European languages. We use the fifth version of the Europarl corpus.
- News Commentary: The News Commentary corpus is from the training data released for the shared tasks of the last few workshops for statistical machine translation (WMT).

Table 3 Size of baseline corpora.

Language Pair	Corpora	Size (lines)
English-Latvian	DGT, JRC	2,305,674
English-Lithuanian	DGT, JRC	2,339,905
English-Estonian	DGT, JRC	2,239,791
English-Slovenian	DGT, JRC	2,190,704
German-Romanian	DGT, JRC	615,336
Latvian-Lithuanian	DGT, JRC	974,161
Lithuanian-Romanian	DGT, JRC	940,461
English-Greek	SETimes	169,337
English-Croatian	SETimes	157,950
English-Romanian	SETimes	171,573

Language Pair	Corpora	Size (lines)
Greek-Romanian	SETimes	175,019
German-English	Europarl, Newscommentary	1,639,893

JRC and DGT are both based on the AC, but they are not identical as they are collected in different ways. JRC consists of more documents than DGT. Whereas the JRC is comprised of mostly full texts, DGT is a collection of translation units, namely segments of translations. The full texts of AC cannot be reproduced from DGT. Most parts of DGT have been manually corrected, while the documents in JRC were aligned automatically without manual validation. Hence, we included both corpora for our SMT baselines.

Table 3 shows the size of the training data we used to train the baseline systems for all the translation directions. We conducted three groups of directions in our experiments. The first group uses JRC and DGT for training and the second group uses SETimes. Although the data combining JRC and DGT is fairly large in size, the domain of the data is rather limited to legislation/law. The systems based on such a data set performed poorly on general translation tasks of other open domains in spite of the high translation quality for in-domain tests reported in previous literature. Therefore, we still consider these language pairs under-resourced. The second group is the opposite. This group of baseline systems is based on the SETimes corpus, which covers a relatively broad range of topics while the size is much smaller than JRC or DGT. The third group includes only German→English as a control group. We used both Europarl and News Commentary for this group. This data set has a presumably open domain and large size. This setup allows us to have more contrastive studies on the effect of using comparable corpora, as the set up for German→English has been used for state of the art systems.

As for language model training, we only use the target portion of the corresponding parallel data. No additional monolingual data is included in our baseline systems.

2.2. Comparable Corpus

To enrich the baseline SMT systems, we use data extracted from comparable corpora collected in the ACCURAT project. We distinguish between the data extracted from news data (USFD-News) and USFD-Wikipedia articles (USFD-Wiki). Details about the collection of these corpora can be found in D3.4, and statistics are reported in D3.6.

2.2.1. Parallel Data

The partners of the ACCURAT consortium used the ACCURAT toolkit to extract parallel sentences from the aligned comparable corpora. D2.6 reports on the particulars of this approach, especially the LEXACC tool. Table 4 gives the statistics about the extracted data. We see that the amount of data varies a lot between language pairs and also the two comparable corpora. We will discuss in Section 5 how this influences translation quality.

Table 4 Statistics of the extracted parallel data.

Language Pair	Corpora	Size (lines)
English-Latvian	USFD-News	112,398
	USFD-Wiki	116,240

Language Pair	Corpora	Size (lines)
English-Lithuanian	USFD-News	33,219
	USFD-Wiki	179,578
English-Estonian	USFD-News	19,048
	USFD-Wiki	128,939
English-Slovenian	USFD-News	67,508
	USFD-Wiki	5,418
German-Romanian	USFD-News	10,227
Latvian-Lithuanian	USFD-News	7,163
	USFD-Wiki	29,370
Lithuanian-Romanian	USFD-News	9,470
English-Greek	USFD-News	6,641
	USFD-Wiki	45,646
English-Croatian	USFD-News	36,663
	USFD-Wiki	31,048
English-Romanian	USFD-News	23,820
	USFD-Wiki	45,771
Greek-Romanian	USFD-News	1,783
German-English	USFD-News	13,782

2.2.2. Monolingual Data

We also want to use the comparable corpora to adapt the language models, but the amount of extracted data is too small to be useful. Instead we make use of the entire USFD-News corpus that was collected in the respective target language. This leads to the amount of data reported in Table 5.

Table 5 Statistics about monolingual comparable corpora.

Language	Size (lines)
Croatian	180,908
German	1,485,774
Greek	1,267,731
English	2,235,282
Estonian	711,147
Latvian	789,178

Language	Size (lines)
Lithuanian	1,061,713
Romanian	1,815,170
Slovenian	470,782

2.3. Development Data

We tune all models on the same development set to get comparable results. The tuning is performed using Minimal Error Rate Training (MERT).

Additionally we make use of the target language tuning texts to interpolate the language models as described in Section 1.3.1.

Table 6 Statistics about development data.

Language Pair	Name of Development Set	Length (in lines)
English→Latvian	Tilde	1000
English→Lithuanian	Tilde	1000
English→Estonian	Tilde	1000
English→Greek	SETimes	600
English→Croatian	SETimes	600
Croatian→English	SETimes	600
English→Romanian	SETimes	600
Romanian→English	SETimes	600
English→Slovenian	mtserver	1000
Slovenian→English	mtserver	1000
German→English	WMT-dev 2008	2051
German→Romanian	RACAI	3000
Romanian→German	RACAI	3000
Greek→Romanian	SETimes	600
Romanian→Greek	SETimes	600
Lithuanian→Romanian	DGT-dev	3000
Latvian→Lithuanian	Tilde	1000

2.4. Test Data

To be able to compare the translation results, we evaluate all systems using the test-balanced test data. This test set contains 511 sentences for all ACCURAT languages in the general domain.

3. Experiments

In this section we describe our experimental setup. Section 3.1 lists all the language pairs we worked on and describes some of the problems SMT has to deal with when translating into/from these languages. Then we explain how we trained the baseline systems and enriched systems. We will discuss the results of our experiments in Section 4.

3.1. Language Pairs

In total, we worked on twelve language pairs. For five of these, we investigated both translation directions. We can group the languages into several groups based on the language family they belong to.

1. Balto-Slavic languages:
    a. Latvian
    b. Lithuanian
    c. Slovenian
    d. Croatian
2. Uralic languages:
    a. Estonian
3. Hellenic languages:
    a. Greek
4. Romance:
    a. Romanian
5. Germanic:
    a. German
    b. English

These languages differ a lot in their different grammatical features. For example, Latvian knows seven different cases, whereas English has none. In our experiments, we examine the following translation directions:

- English→Latvian
- English→Lithuanian
- English→Estonian
- English→Greek
- English→Croatian
- Croatian→English
- English→Romanian
- Romanian→English
- English→Slovenian
- Slovenian→English
- German→English
- German→Romanian

- Romanian→German
- Greek→Romanian
- Romanian→Greek
- Lithuanian→Romanian
- Latvian→Lithuanian

Our main concern is to translate from English, but we also investigate a few language pairs that do not involve English for which there is very little data available.

3.2. Baseline Systems

We retrained the baseline systems listed in D4.1, as we wanted to also include the interpolated language models for the baseline models.

3.2.1. Training Method

We trained state-of-the-art phrase-based models using 7-gram phrase-tables and 5-gram interpolated language models. For the training we used the data described in Section 2.1, where the parallel data was used for the translation model and the target language text to generate the language model. In the case of the language pairs using DGT and JRC as well as German→English, we interpolated the language models built on the two baseline corpora using the target side of our development set. This is the same set that we later optimised the SMT translation parameters on using Minimal Error Rate Training (MERT) and is listed in Table 6.

3.3. Enriched SMT Systems

For each of the baseline systems, we trained systems using the additional data described in Section 2.2. We train separate models for the data extracted from the USFD-News and the USFD-Wiki data to examine the influence the different sorts of data has: whereas USFD-News consists of text that have been aligned automatically using document alignment software such as DictMetric (see D1.3 for details on DictMetric, and D2.2 for details on the alignment process), USFD-Wiki corpus includes an inherent alignment that can be created by using the inter-Wikipedia links between articles describing the same subject in different languages. Hence, document alignment is much easier and less error prone for the USFD-Wiki corpus. On the other hand, in USFD-News we can have many-to-many document alignments, which might contain alternate translations for the same input sentences, thus increasing the translation options in our phrase table.

3.3.1. Data

Table 7 lists the amount of training data used for each language pair.

Table 7 Statistics of training data for enriched SMT systems.

Language Pair	Parallel Corpora	Size (lines)	Monolingual Corpora	Size (lines)
English→Latvian	DGT, JRC, USFD-News	2,418,072	DGT, JRC, USFD-News	3,094,852

Language Pair	Parallel Corpora	Size (lines)	Monolingual Corpora	Size (lines)
	DGT, JRC, USFD-Wiki	2,421,914		
English→Lithuanian	DGT, JRC, USFD-News	2,373,124	DGT, JRC, USFD-News	3,401,618
	DGT, JRC, USFD-Wiki	2,519,483		
English→Estonian	DGT, JRC, USFD-News	2,258,839	DGT, JRC, USFD-News	2,950,938
	DGT, JRC, USFD-Wiki	2,368,730		
English→Greek	SETimes, USFD-News	175,978	SETimes, USFD-News	1,437,068
	SETimes, USFD-Wiki	214,983		
English→Croatian	SETimes, USFD-News	194,613	SETimes, USFD-News	338,858
	SETimes, USFD-Wiki	188,998		
Croatian→English	SETimes, USFD-News	194,613	SETimes, USFD-News	2,393,232
	SETimes, USFD-Wiki	188,998		
English→Romanian	SETimes, USFD-News	195,393	SETimes, USFD-News	1,986,743
	SETimes, USFD-Wiki	217,344		
Romanian→English	SETimes, USFD-News	195,393	SETimes, USFD-News	2,406,855
	SETimes, USFD-Wiki	217,344		
English→Slovenian	DGT, JRC, USFD-News	2,258,212	DGT, JRC, USFD-News	2,661,486
	DGT, JRC, USFD-Wiki	2,196,122		
Slovenian→English	DGT, JRC, USFD-News	2,258,212	DGT, JRC, USFD-News	4,425,986
	DGT, JRC,	2,196,122		

Language Pair	Parallel Corpora	Size (lines)	Monolingual Corpora	Size (lines)
	USFD-Wiki			
German→English	Europarl, NC, USFD-News	1,653,675	Europarl, NC, USFD-News	3,875,175
German→Romanian	DGT, JRC, USFD-News	625,563	DGT, JRC, USFD-News	2,430,506
Romanian→German	DGT, JRC, USFD-News	625,563	DGT, JRC, USFD-News	2,101,110
Greek→Romanian	SETimes, USFD-News	176,802	SETimes, USFD-News	1,990,189
Romanian→Greek	SETimes, USFD-News	176,802	SETimes, USFD-News	1,442,750
Lithuanian→Romanian	DGT, JRC, USFD-News	949,931	DGT, JRC, USFD-News	2,655,631
Latvian→Lithuanian	DGT, JRC, USFD-News	981,324	DGT, JRC, USFD-News	2,035,874
	DGT, JRC, USFD-Wiki	1,003,891		

3.3.2. Training Method

We use the same general settings for training the enriched models as we did for training the baseline models. In this task we want to focus on the influence the additional data that we have.

3.3.2.1. Interpolated LM

For the interpolated language model, we use the target side of both the baseline parallel data as well as the collected comparable corpus. The translation model is trained on the extracted parallel data and the baseline corpora. We apply this approach to both the USFD-News and the USFD-Wiki data.

The extracted data is very small, though (cf. Section 2.2, Table 4). Instead we use the entire comparable corpus, as this provides us with much more data. As the language model only deals with the target language, the monolingual corpus can be anything and does not need to be part of the parallel corpus. We use the comparable News corpus to train the language model for both the USFD-News and USFD-Wiki experiments.

We use the same development sets as for the baseline systems. The target language texts of those sets are then also used during the interpolation of the language models.

3.3.2.2. Mixture Model

Following the method described in Section 1.2.2, we train phrase table on each individual corpus and then combine them into a single mixture translation model. For the language model, we use the interpolated language models from the systems described in Section 3.3.2.1. For these experiments, we only used the USFD-News corpora.

The systems are tuned using the same development sets as before.

4. Results

In this section we present the results of our experiments. All systems were tested on the same test set, *test-balanced*, consisting of 511 sentences. This test set contains general domain text.

This section is organised as follows: Section 4.1 will report on automatic scores such as BLEU. We also investigated the influence of the LEXACC score on translation quality as measured by BLEU and report on the results of these experiments in Section 4.2.

4.1. BLEU Scores

Since we retrained the baseline systems as described in Section 3.2, we also had to regenerate the BLEU scores for these systems. Table 8 lists the results for all experiments on interpolated language models and mixture models. Figures in bold indicate models that outperform the baseline. The best model for each language pair is denoted with an asterisk.

Table 8 BLEU scores for all experiments.

Language Pair	Baseline	Interpolated LM		Mixture Models
		USFD-News	USFD-Wiki	
English→Latvian	12.74	**13.20** (+.46)	**13.07** (+.33)	**13.25*** (+.51)
English→Lithuanian	12.66	12.21 (-.45)	12.33 (-.33)	11.94 (-.71)
English→Estonian	10.44	**11.23*** (+.79)	**10.46** (+.02)	**10.88** (+.44)
English→Greek	19.06	**21.40** (+2.34)	**23.67*** (+4.61)	**20.61** (+1.55)
English→Croatian	10.91	10.36 (-.55)	**11.25** (+.34)	**11.45*** (+.54)
Croatian→English	20.78	20.31 (-.47)	**21.17** (+.39)	**21.91*** (+1.13)
English→Romanian	17.89	**20.11*** (+2.22)	**20.00** (+2.11)	**19.08** (+1.19)
Romanian→English	21.54	**26.16** (+4.62)	**30.35*** (+8.81)	**25.27** (+3.73)
English→Slovenian	18.20	**18.68*** (+.48)	**18.66** (+.46)	17.70 (-.50)
Slovenian→English	26.28	**27.40** (+1.12)	**27.46*** (+1.18)	**27.31** (+1.03)
German→English	27.90	**28.62*** (+.72)	#	27.88 (-.02)
German→Romanian	9.66	**10.14*** (+.48)	#	8.37 (-1.29)
Romanian→German	10.22	9.56 (-.66)	#	9.97 (-.25)
Greek→Romanian	15.81	**17.25*** (+1.44)	#	**17.15** (+1.34)
Romanian→Greek	12.13	**13.59*** (+1.46)	#	**13.37** (+1.24)
Lithuanian→Romanian	9.91	9.24 (-.67)	#	4.67 (-5.24)
Latvian→Lithuanian	12.12	**12.69*** (+.57)	8.70 (-3.42)	**12.41** (+.29)

We see that not every approach works equally well for each language direction. For some of the language pairs we don't observe any improvement by adding the data, thus we investigated English→Lithuanian, Romanian→German and Lithuanian→Romanian further. We describe these experiments in the next section.

The largest improvement in BLEU score can be noted for those language pairs which only used the SETimes corpus with less than 200,000 lines per language pair as baseline corpus. For the language pairs using DGT/JRC the improvements are smaller.

4.2. Out Of Vocabulary Counts

As we observe degradations for some of the language pairs, it is worthwhile to note the *out of vocabulary* (OOV) counts for each model. This count represents how many tokens in the test data are not contained in the training data and thus cannot be translated properly. By adding more data, we hope to decrease the OOV count and receive better translation. Table 9 gives the OOV counts for all models. The counts are given (in per cent out of the total tokens in the test set) for tokens (counting all unknown tokens) and types (only counting unique unknown tokens). Languages that performed badly are marked with asterisks.

Table 9 OOV counts for all MT models for test-balanced test set.

Language Pair	Corpus	OOV Source	OOV Reference
English→Latvian	Baseline	4.2% / 1.1%	8.8% / 3.8%
	USFD-News	2.0% / 0.5%	4.8% / 2.1%
	USFD-Wiki	3.5% / 0.9%	8.5% / 3.7%
English→Lithuanian*	Baseline	4.1% / 1.1%	9.1% / 4.1%
	USFD-News	2.1% / 0.5%	6.5% / 2.9%
	USFD-Wiki	1.4% / 0.3%	5.1% / 2.3%
English→Estonian	Baseline	4.1% / 1.1%	14.5% / 7.2%
	USFD-News	2.6% / 0.7%	12.7% / 6.3%
	USFD-Wiki	1.5% / 0.4%	10.3% / 5.2%
English→Greek	Baseline	6.0% / 1.8%	13.1% / 4.8%
	USFD-News	5.5% / 1.5%	12.4% / 4.5%
	USFD-Wiki	3.1% / 0.8%	8.2% / 2.9%
English→Croatian	Baseline	6.0% / 1.9%	17.2% / 8.1%
	USFD-News	4.1% / 1.2%	13.0% / 5.9%
	USFD-Wiki	4.3% / 1.2%	14.0% / 6.3%
Croatian→English	Baseline	17.2% / 8.1%	6.0% / 1.9%
	USFD-News	13.0% / 5.9%	4.1% / 1.2%
	USFD-Wiki	14.0% / 6.3%	4.3% / 1.2%
English→Romanian	Baseline	6.6% / 2.0%	24.3% / 14%
	USFD-News	5.3% / 1.4%	17.1% / 6.8%

Language Pair	Corpus	OOV Source	OOV Reference
	USFD-Wiki	3.5% / 0.9%	9.0% / 3.3%
Romanian→English	Baseline	24.3% / 14%	6.6% / 2.0%
	USFD-News	17.1% / 6.8%	5.3% / 1.4%
	USFD-Wiki	9.0% / 3.3%	3.5% / 0.9%
English→Slovenian	Baseline	4.4% / 1.1%	7.8% / 3.2%
	USFD-News	2.2% / 0.5%	4.4% / 1.8%
	USFD-Wiki	4.0% / 1.0%	7.3% / 3.0%
Slovenian→English	Baseline	7.8% / 3.2%	4.4% / 1.1%
	USFD-News	4.4% / 1.8%	2.2% / 0.5%
	USFD-Wiki	7.3% / 3.0%	4.0% / 1.0%
German→English	Baseline	7.5% / 4.0%	2.2% / 1.3%
	USFD-News	7.3% / 2.4%	2.0% / 0.5%
German→Romanian	Baseline	18.0% / 6.4%	12.2% / 4.4%
	USFD-News	13.4% / 4.4%	8.9% / 3.1%
Romanian→German*	Baseline	12.2% / 4.4%	18.0% / 6.4%
	USFD-News	8.9% / 3.1%	13.4% / 4.4%
Greek→Romanian	Baseline	13.2% / 4.8%	24.3% / 14.0%
	USFD-News	13.1% / 4.7%	24.0% / 10.7%
Romanian→Greek	Baseline	24.3% / 14.0%	13.2% / 4.8%
	USFD-News	24.0% / 10.7%	13.1% / 4.7%
Lithuanian→Romanian*	Baseline	11.6% / 5.2%	10.6% / 3.6%
	USFD-News	11.6% / 5.2%	7.9% / 2.7%
Latvian→Lithuanian	Baseline	7.6% / 3.3%	9.1% / 4.1%
	USFD-News	7.6% / 3.3%	8.0% / 3.6%
	USFD-Wiki	7.1% / 3.4%	7.5% / 3.4%

As we see, the counts decrease in varying degrees for USFD-News and USFD-Wiki.

Although the OOV counts decrease for most languages, in Lithuanian→Romanian, one of the language pairs that performed badly, we see that the OOV count for the source side of our test set remains stable, i.e. no unknown words from Lithuanian are covered by the additional data. We see a decrease for the Romanian side, but this does not affect translation, as we won't be able to match them to the corresponding source words.

Additionally we can observe that the biggest improvements in BLEU correspond to those languages with the largest decrease in OOV counts, such as Romanian→English.

The OOV count accounts for why the experiments for Lithuanian→Romanian do not achieve an improvement in BLEU score. It does not explain why English→Lithuanian and Romanian→German perform so badly, as we see a large decrease in OOV counts for both language pairs.

4.3. Staggered Experiments

The LEXACC tool assigns each sentence pair a score that denotes how likely these two sentences are parallel. As such, the LEXACC score should allow us to predict how usable a particular chunk of the data is, i.e. that using this data will increase translation quality.

To test this influence of the LEXACC score, we split up the extracted data. We want to check the effect of the score both in intervals and in a cumulative fashion. The hypothesis for the former is that data with a higher LEXACC score should help more than data with a lower score. In the cumulative experiments we choose different thresholds. As the score goes down, the less parallel the data will become and more errors will be introduced into the translation model. But as the distribution of the data follows Zipf's law, we have very few items with a very high score, but the lower the score, the more sentences LEXACC extracts. But we also need to take into account how much data we have, i.e. for higher thresholds LEXACC will only be able to extract small amounts of data. Here we are interested in the threshold that allows the maximal increase in translation quality for the amount of data used. This threshold may vary for different corpora, an effect we also want to examine.

As we couldn't observe an improvement in translation quality in the experiments using the full data for English→Lithuanian, Romanian→German and Lithuanian→Romanian, we treat these languages in these experiments. Additionally we examine English→Latvian and English→Romanian. In these two languages we saw improvements, but we are interested in seeing how much each part of the data contributes. We chose them because they work with different baseline corpora, so we can see the effects of adding a small amount of data to a large out-of-domain corpus (DGT/JRC in the case of English→Latvian) and adding similar amounts of data to a small in-domain corpus (SETimes for English→Romanian).

4.3.1. English→Latvian

For English→Latvian we examined both the interpolated language models as well as the mixture models. The problem with using mixture models is that on such a small set of data the probabilities associated with the entries in the phrase table become less trustworthy. Table 10 and Table 11 give the amount of data (in sentence pairs) in the different intervals.

Table 10 Statistics about interval experiments for English->Latvian.

Interval	USFD-News	USFD-Wiki
>0.9	169	208
0.9 – 0.8	3226	1730
0.8 – 0.7	13264	5791
0.7 – 0.6	12735	6868
0.6 – 0.5	9009	7085
0.5 – 0.4	6914	8556
0.4 – 0.3	8720	13902

Interval	USFD-News	USFD-Wiki
0.3 – 0.2	15325	26669
0.2 – 0.1	43036	45431

We did not investigate data with a LEXACC score of less than 0.1 (the default threshold of LEXACC is 0.1). We see that we have very little data with a score higher than 0.9, but for lower scores we get more data.

Table 11 Statistics about cumulative experiments for English->Latvian.

Cumulative	USFD-News	USFD-Wiki
>0.9	169	208
>0.8	3395	1938
>0.7	16659	7729
>0.6	29394	14597
>0.5	38403	21682
>0.4	45317	30238
>0.3	54037	44140
>0.2	69362	70809
>0.1	112398	116240

We used each chunk of the data to retrain the SMT model and evaluated it the same as the baseline and full enriched models. Table 12 and Table 13 give the BLEU scores for those experiments. The baseline SMT system reached a BLEU score of 12.66. Experiments that perform worse than the baseline are marked in italic; the best experiment in each approach and corpus is marked in boldface.

Table 12 BLEU scores for interval experiments for English->Latvian.

Interval	Interpolated LM		Mixture Models	
	USFD-News	USFD-Wiki	USFD-News	USFD-Wiki
>0.9	13.48	13.73	12.97	13.48
0.9 – 0.8	13.60	13.57	**13.29**	13.36
0.8 – 0.7	13.15	13.57	12.71	13.29
0.7 – 0.6	**13.67**	**13.83**	12.76	13.23
0.6 – 0.5	13.49	13.50	12.84	12.91
0.5 – 0.4	13.54	13.57	12.78	**13.72**
0.4 – 0.3	13.31	13.39	12.80	13.61
0.3 – 0.2	12.77	13.40	12.99	13.44

Interval	Interpolated LM		Mixture Models	
	USFD-News	USFD-Wiki	USFD-News	USFD-Wiki
0.2 – 0.1	*12.15*	*12.63*	12.84	12,86

Figure 1 illustrates the effect of the LEXACC score on the BLEU score. The data in the interval of [0.1,0.2] scores the worst results and doesn't even reach the BLEU score of the baseline (plotted for comparison purposes). As the LEXACC score increases, we can also see an increase in BLEU score. Using the interpolated language models, this development is rather steady. When we compare USFD-News to the USFD-Wiki extracted data, the interpolated language models show similar trends.

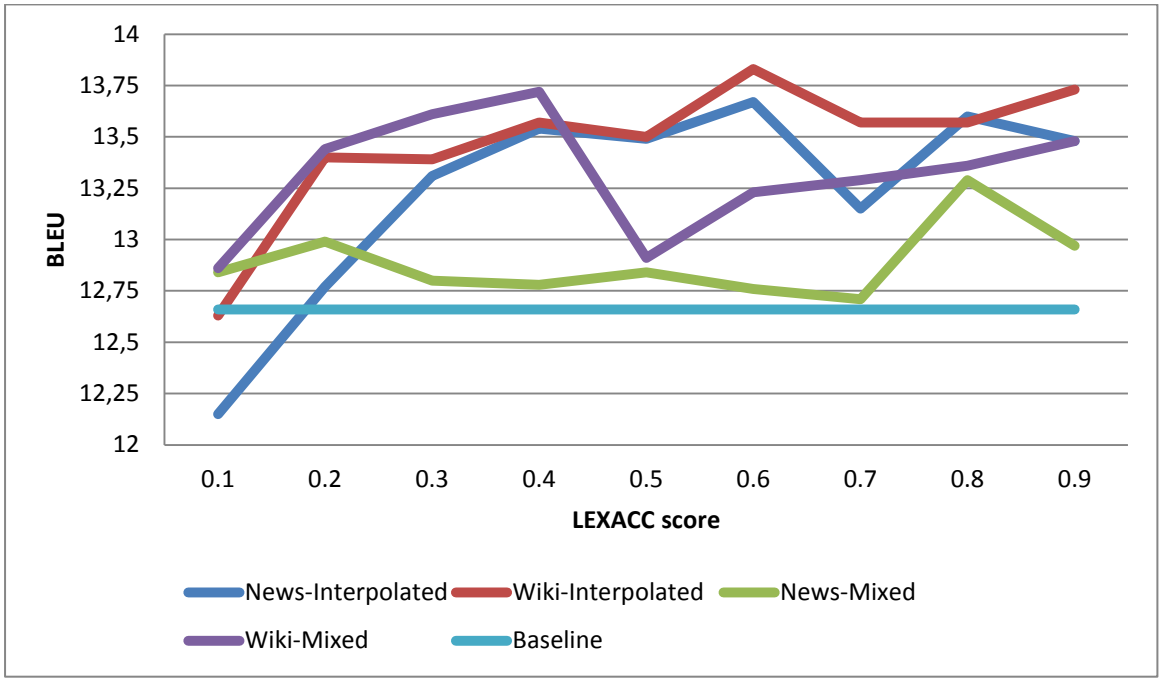

Figure 1 BLEU scores for interval experiments for English->Latvian.

According to the BLEU scores, the translation results using the mixture models seem less correlated to the LEXACC score, mostly due to the fact that the mixture models is very sensitive to the size of the data that is used to construct the additional phrase tables. Higher LEXACC thresholds indicate better quality of extracted sentence pairs. Meanwhile, these high scores also result in less extracted data. The translation model constructed over a small amount of data tend to contain less useful phrase pair entries while having high probability estimation values in general. When combining a small model with high scores with a much larger model with much lower scores, it is not avoidable to penalize the phrase pairs from the small model in order to use entries that exist in the other models, which are actually the majority of the combined model. Thus, the tuning procedure seems to assign in general higher weights to the feature that represents the larger model. As a result, the additional data could not have as much influence on the final translation as we hope. It also explains why in the experiment for USFD-Wiki data the BLEU score drops significantly at the LEXACC interval [0.4,0.5], for which there are nearly 40% less sentence pairs than for [0.3,0.5]. The BLEU score increases again for higher LEXACC scores, as the size difference is smaller for

the other cases. In practice, the probability estimation in the sub-models should all be normalized, but this would make it more difficult to compare results for different extracted data. Therefore, we chose to retain the probability scores in the sub-models.

The results for the cumulative experiments are not quite as clear. The effect of the LEXACC score on BLEU is plotted in Figure 2. Here we see a lot of fluctuation. Although the best BLEU scores are comparable for three of the four experiment runs, they occur in different intervals. Especially interesting is the behaviour of the data with a LEXACC score of 0.7 and above. In USFD-News this chunk leads to an improvement using the interpolated LMs, but for the mixture models the BLEU score drops by almost 0.6, a significant deterioration. The USFD-Wiki data behaves similarly, except that here the BLEU score of the interpolated LM drops even underneath the baseline performance, whereas this data is the best performing for the mixture models.

Table 13 BLEU scores for cumulative experiments for English->Latvian.

Cumulative	Interpolated LM		Mixture Models	
	USFD-News	USFD-Wiki	USFD-News	USFD-Wiki
>0.9	13.48	**13.73**	12.97	13.48
>0.8	13.50	13.34	13.77	12.90
>0.7	13.66	*12.56*	13.19	**13.49**
>0.6	**13.86**	13.55	**13.78**	12.97
>0.5	13.73	13.10	13.00	13.11
>0.4	13.68	13.30	13.41	12.90
>0.3	13.58	13.22	13.26	12.96
>0.2	13.74	13.46	13.75	13.15
>0.1	13.20	13.07	13.25	#

Figure 2 illustrates this point. We see a lot of ups and downs, although the data using a threshold of 0.6 seems to work reliably well for both models and both corpora.

One of the questions we wanted to investigate with our experiments was if the LEXACC score correlates to the BLEU score. To answer this question, we also examine the evaluation results of LEXACC alone Table 14 lists the performance of LEXACC on a test corpus in which each parallel pair was diluted by 100 noisy pairs, i.e. sentences that were not translations of one another. Using the gold standard parallel sentences, the results of LEXACC were evaluated using the precision and recall metrics as well as the F1 score. We see that for English→Latvian, LEXACC performs best for a threshold of 0.52. In our experiments using data from this interval does not perform best, but the BLEU scores are close to the best performing model for the interpolating language models, whereas the mixture models are more sensitive to the quality of input.

Table 14 LEXACC6 performance scores on the 100to1 corpus with document alignments.

Language Pair	F1	P	R	Threshold
English→German	65.91	76.32	58.00	0.32
English→Greek	75.45	94.03	63.00	0.31
English→Estonian	53.59	77.36	41.00	0.20
English→Lithuanian	64.05	92.45	49.00	0.30
English→Latvian	75.58	90.28	65.00	0.52
English→Romanian	56.60	76.27	45.00	0.36
English→Slovenian	34.33	67.65	23.00	0.29

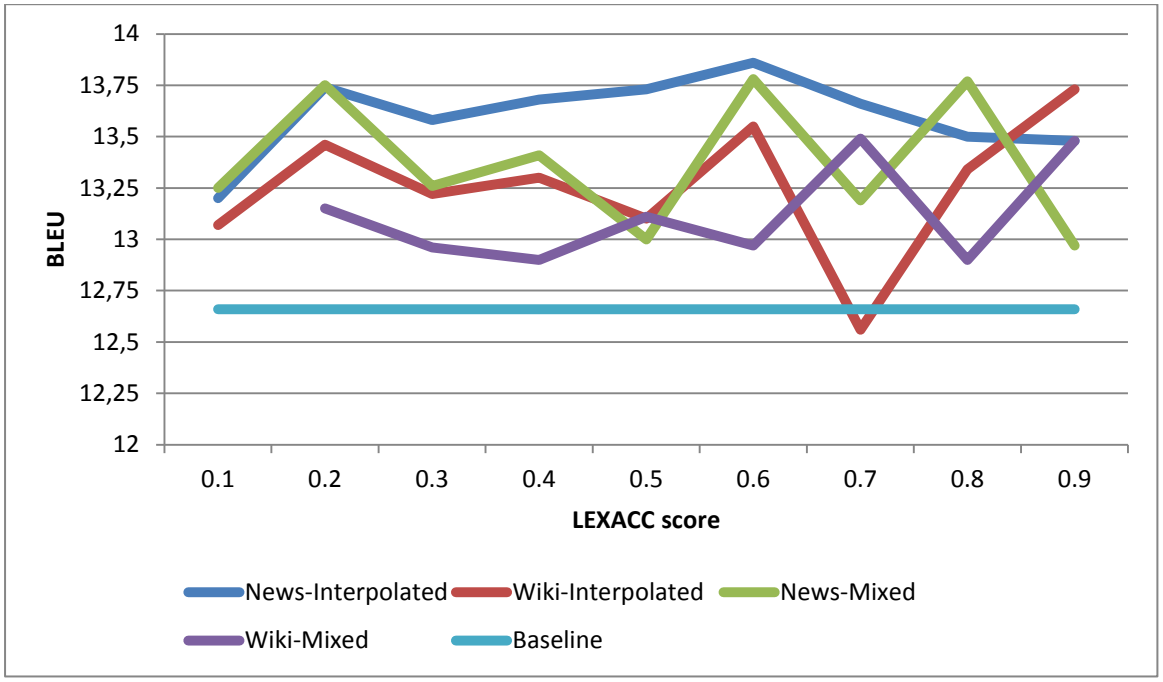

Figure 2 BLEU scores for cumulative experiments for English->Latvian.

4.3.2. English→Romanian

The training data for English→Romanian was very small, so our hypothesis was that this language direction was very sensitive to the quality of the newly added data. Whereas the DGT/JRC corpora are big enough to smooth out mistakes in the translation probabilities, the SETimes corpus is small enough that even the relatively small amount of extracted data can counteract the probabilities extracted from the original data: the English-Latvian baseline corpus consists of 2,305,674 lines, with 112,398/116,240 lines extracted from each comparable corpus, adding about 5% of data to the baseline corpus. For English-Romanian, we only had 171,573 lines in the baseline, so the data from USFD-News (23,8320 lines) and the USFD-Wiki corpus (45,771 lines) amount to 14% and 27% respectively. Thus the influence of the new data will be much higher than for the previous experiments.

For this language pair we examined only the interpolated language models as the results on the mixture models were too unsteady. Table 15 and Table 16 give the amount of data in the different intervals.

Table 15 Statistics about interval experiments for English->Romanian.

Interval	USFD-News	USFD-Wiki
>0.9	246	5807
0.9 – 0.8	2468	13174
0.8 – 0.7	2221	6530
0.7 – 0.6	1511	3993
0.6 – 0.5	2021	3653
0.5 – 0.4	2636	3974
0.4 – 0.3	4024	3826
0.3 – 0.2	8693	4814

The distribution of this data is especially interesting. In English→Latvian the distribution followed Zipf's law, i.e. there was very little data for the high scores, but the lower the score the more data was extracted. For English-Romanian, however, this only holds for USFD-News. The USFD-Wiki corpus behaves differently: here we have unusually many sentence pairs with a high score. This cannot be simply explained with the fact that USFD-Wiki articles are inherently more strongly comparable than news text, as then this would also have to hold for other language pairs. Manual inspection of the data suggests that many articles in the Romanian Wikipedia have been originally translated from the English Wikipedia. We consider this an anomaly.

We did not investigate data with a LEXACC score of less than 0.2.

Table 16 Statistics about cumulative experiments for English->Romanian.

Cumulative	USFD-News	USFD-Wiki
>0.9	246	5807
>0.8	2714	18981
>0.7	4935	25511
>0.6	6446	29504
>0.5	8467	33157
>0.4	11103	37131
>0.3	15127	40957
>0.2	23820	45771

The procedure of these experiments is the same as for the previous English→Latvian experiments. For each chunk of the data, we retrain the SMT models and compare it against the baseline, which was evaluated with a BLEU score of 17.89.

Table 17 BLEU scores for interval experiments for English->Romanian.

Interval	Interpolated LM	
	USFD-News	USFD-Wiki
>0.9	19.45	19.08
0.9 – 0.8	18.74	19.63
0.8 – 0.7	19.28	**19.64**
0.7 – 0.6	19.81	18.79
0.6 – 0.5	20.03	19.13
0.5 – 0.4	20.04	19.29
0.4 – 0.3	**20.22**	19.30
0.3 – 0.2	19.92	18.30

All systems outperform the baseline, but the overall tendency for improvement of BLEU is not as clear-cut as it was for the previous experiment. Instead we see that the improvement in BLEU varies a lot over of the intervals. For the USFD-Wiki corpus, which adds 25% to the original data, our assumption that higher LEXACC scores predict a higher increase in BLEU still holds, but for the USFD-News data we find that using the maximum of available data results in the highest gain. Here we must take into account the amount of data in each interval: although USFD-Wiki can offer us 13,000 additional lines in the interval of [0.9,0.8], there are only 2,500 sentences in the same interval in the News corpus.

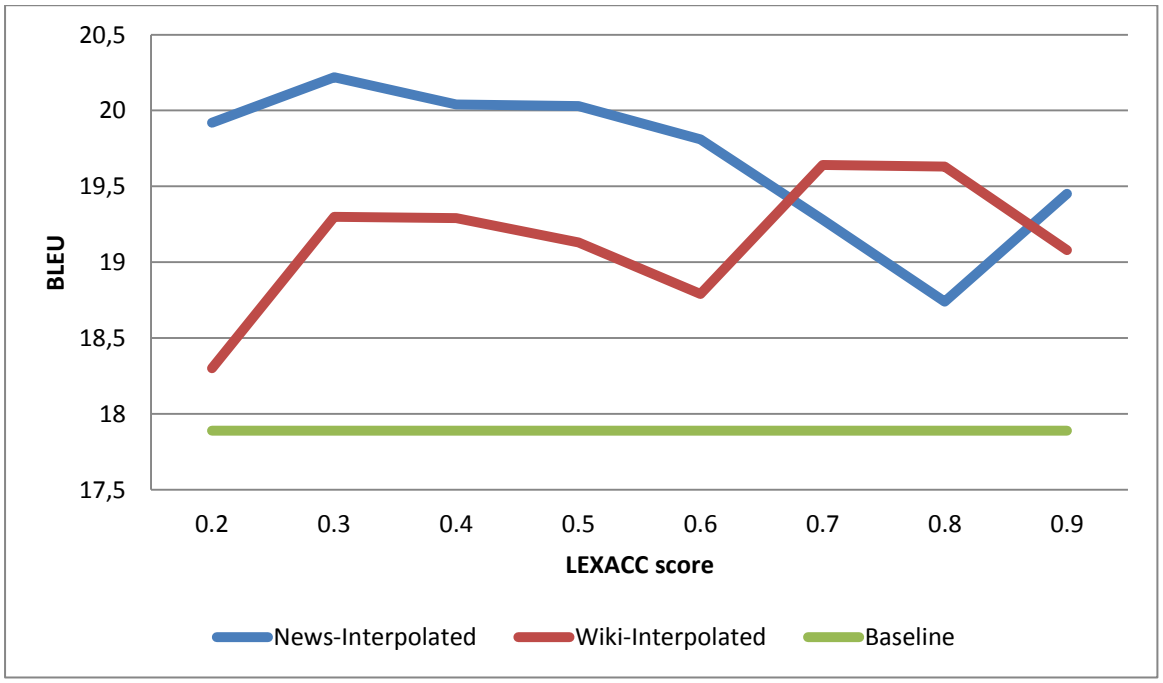

Figure 3 BLEU scores for interval experiments for English->Romanian.

Table 18 shows the results for the cumulative experiments. As for the interval experiments, all models improve over the baseline.

Table 18 BLEU scores for cumulative experiments for English->Romanian.

Cumulative	Interpolated LM	
	USFD-News	USFD-Wiki
>0.9	19.45	19.08
>0.8	19.04	19.59
>0.7	18.54	19.75
>0.6	18.71	20.03
>0.5	19.01	19.98
>0.4	19.85	20.27
>0.3	19.44	**20.40**
>0.2	**20.11**	20.00

In Figure 4 we see less variation than for English→Latvian, with rather obvious thresholds for the corpora. As for the interval experiments, we get the best results by using all of the available additional data for the USFD-News corpus, whereas the threshold for USFD-Wiki lies at 0.3. This is consistent with the best LEXACC performance r, where we reach the best F1 score at a threshold of 0.36. Although these thresholds are close, we see quite a difference between the different corpora: the USFD-News corpus improves by 0.7 BLEU points when using all the data, whereas the performance of the USFD-Wiki corpus drops by 0.3 BLEU points when using the same threshold. The BLEU scores for threshold 0.3 differ by almost one full BLEU score, a very significant difference. This can be explained by taking into account the amount of data (see Table 16): for this interval we have almost three times as many sentences for USFD-Wiki than for USFD-News.

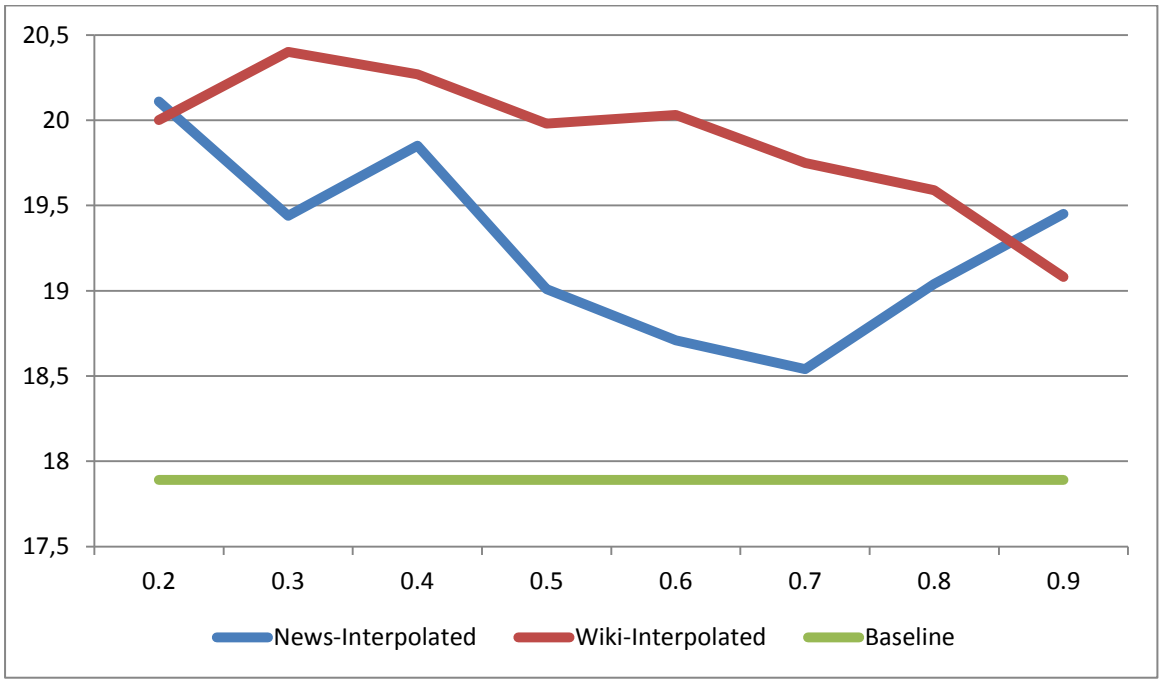

Figure 4 BLEU scores for cumulative experiments for English->Romanian.

4.3.3. English→Lithuanian

As shown in Section 4.1, using the full data did not result in an improvement of BLEU score for English→Lithuanian. As we have seen a lot of variation in the BLEU scores for the individual chunks of the data, we decided to give English→Lithuanian the same treatment, so we could check whether there was, for example, one particularly bad batch of data that affected the performance of the overall system.

The size of the original baseline corpus consisting of DGT/JRC was 2,339,905 lines. To this we could add 33,219 lines from the USFD-News corpus (+1.42%) and 179,578 lines from USFD-Wiki (+7.67%). Splitting up the data into the individual chunks, results in the amount of data shown in Table 19 and Table 20.

Table 19 Statistics about interval experiments for English->Lithuanian.

Interval	USFD-News	USFD-Wiki
>0.9	28	1089
0.9 – 0.8	352	4265
0.8 – 0.7	1006	6450
0.7 – 0.6	1061	6307
0.6 – 0.5	1317	7656
0.5 – 0.4	1692	10393
0.4 – 0.3	2495	17628
0.3 – 0.2	5536	35574
0.2 – 0.1	19732	90196

The data follows again the distribution we would expect. The difference in size between the USFD-News and USFD-Wiki corpus is significant—in each section we have about six times as much data for USFD-Wiki than for the USFD-News corpus. Whereas we can generally explain low extraction numbers by LEXACC with the fact that it heavily depends on the available bilingual lexicons, it is interesting that with the same configuration we get such different amounts of data. For English→Latvian we received comparable amounts of data, although Latvian is similar in the language characteristics as Lithuanian, both belonging to the Balto-Slavic family and they share several features, such as a intricate case system: both have seven cases, for example. Although they differ in the details, extracting Lithuanian text should be no more complicated than extracting Latvian text. As the numbers of the USFD-Wiki corpus for English→Lithuanian are comparable to the English→Latvian numbers, this is an interesting anomaly.

Table 20 Statistics about cumulative experiments for English->Lithuanian.

Cumulative	USFD-News	USFD-Wiki
>0.9	28	1089
>0.8	380	5354
>0.7	1386	11804
>0.6	2447	18111
>0.5	3764	25767
>0.4	5456	36160
>0.3	7951	53788
>0.2	13487	89562
>0.1	33219	179758

The baseline produced a BLEU score of 12.66. Table 21 and Figure 5 present the BLEU scores for the respective interval and cumulative experiments.

Table 21 BLEU scores for interval experiments for English->Lithuanian.

Interval	Interpolated LM	
	USFD-News	USFD-Wiki
>0.9	***12.48***	***12.64***
0.9 – 0.8	*12.00*	*12.49*
0.8 – 0.7	*12.47*	*12.40*
0.7 – 0.6	*12.47*	*12.53*
0.6 – 0.5	*12.33*	*12.37*
0.5 – 0.4	*12.46*	*12.00*
0.4 – 0.3	*12.01*	*12.26*
0.3 – 0.2	*12.04*	*12.34*

Interval	Interpolated LM	
	USFD-News	USFD-Wiki
0.2 – 0.1	*12.13*	*11.87*

None of the interval experiments perform better than the baseline, but we can see that the USFD-Wiki data performs much better than the USFD-News data. We observe in Figure 5 the general tendency that higher scoring intervals result in better BLEU scores, but the amount of data does not seem sufficient to push the enriched system over the baseline.

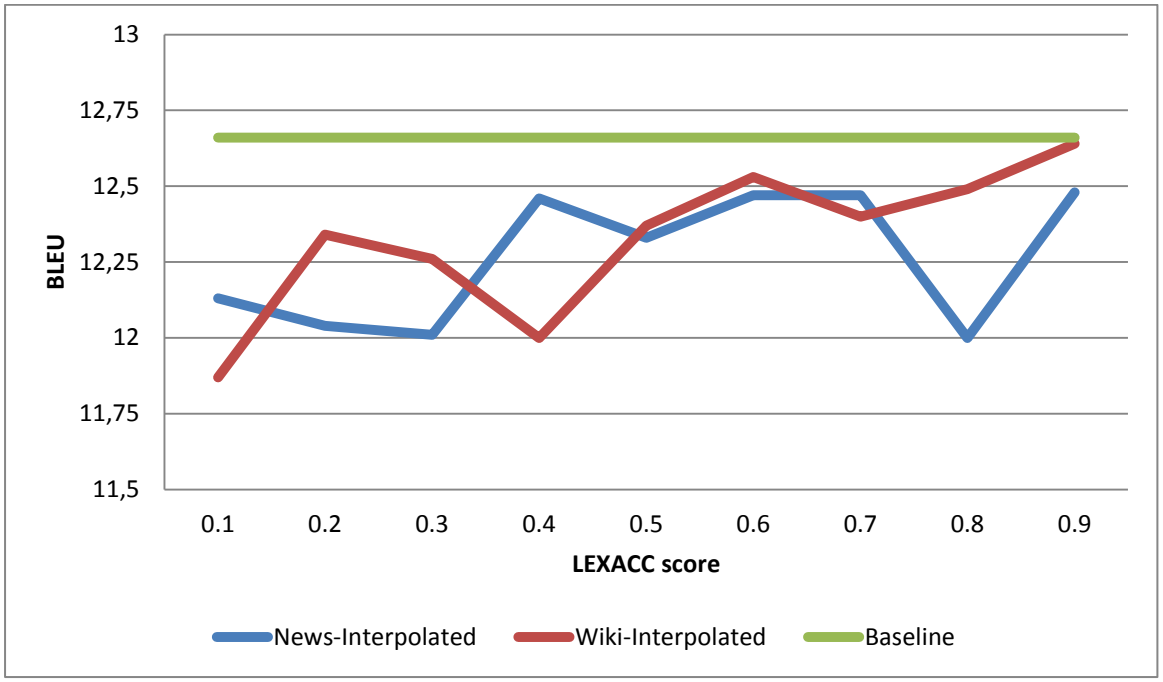

Figure 5 BLEU scores for interval experiments for English->Lithuanian.

Using the interval, especially the small amounts available for the USFD-News corpus did not yield an improvement system.

Table 22 BLEU scores for cumulative experiments for English->Lithuanian.

Cumulative	Interpolated LM	
	USFD-News	USFD-Wiki
>0.9	*12.48*	***12.64***
>0.8	*12.35*	*12.56*
>0.7	*12.35*	*12.34*
>0.6	**12.94**	*12.43*
>0.5	*11.90*	*12.41*
>0.4	*12.11*	*12.32*
>0.3	*12.45*	*12.25*

Cumulative	Interpolated LM	
	USFD-News	USFD-Wiki
>0.2	*12.37*	*11.93*
>0.1	*11.21*	*12.33*

Most of the cumulative experiments also perform worse than the baseline. It is interesting to note that the best-performing system, which also improves over the baseline, uses the same threshold we have already identified as optimal for English→Latvian, namely 0.6. This can be interpreted such that Lithuanian generally behaves similar to Latvian.

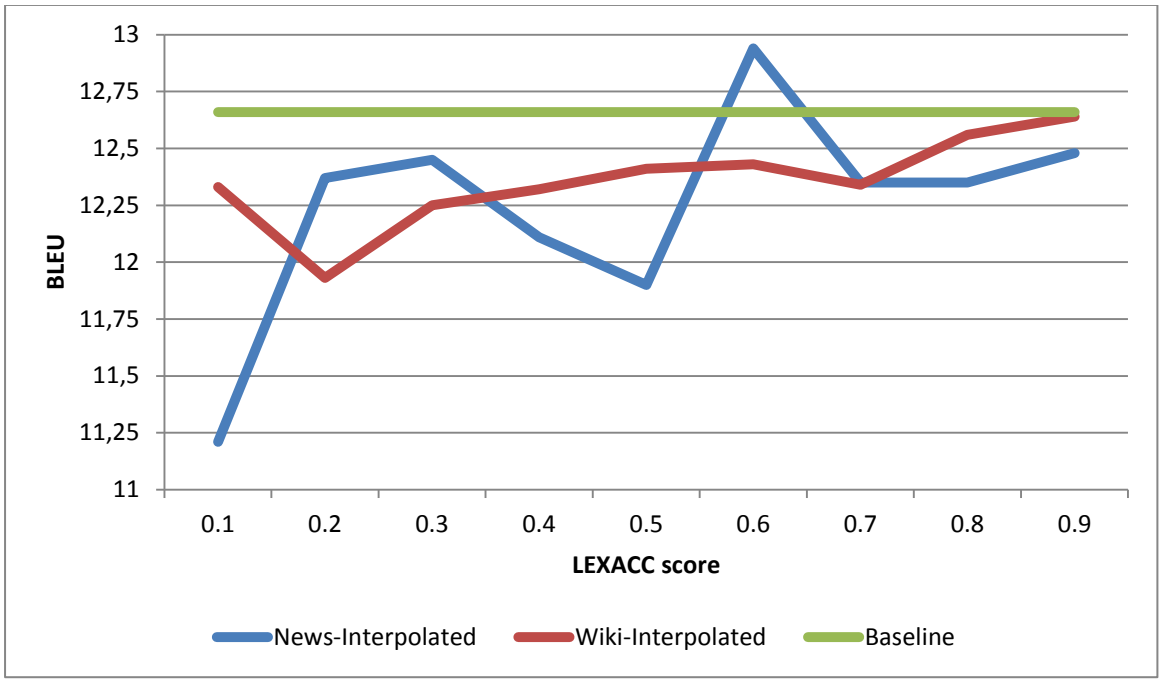

Figure 6 BLEU scores for cumulative experiments for English->Lithuanian.

It is worthwhile to note that the upper intervals get close to the performance of the baseline, which leads us to believe that the amount of data extracted was simply too small to have a large enough impact on the baseline corpus.

When we consider the LEXACC threshold with the highest F1 score from Table 14, we see that there is no correlation between this score (0.3) and the highest BLEU scores (at thresholds 0.6 and 0.9 for the USFD-News and USFD-Wiki corpus respectively).

4.3.4. Lithuanian→Romanian

We have seen in Section 4.2 that the OOV counts already explain why we don't see an improvement. We still would like to see if there's an optimal split in the data so that we can still see an improvement, as for Lithuanian→Romanian we observe the worst degradation, performing 0.57 respectively 5.24 BLEU points worse than the baseline model in both experiments.

When we look at the data we extracted in Table 4, it is immediately obvious that it's only very small amount. The 9,470 extracted sentences correspond to only 1% of the original

baseline DGT/JRC data. In the following experiments we investigate if this is enough to have an impact on translation quality.

We observe that there is no data with scores in the interval [1.0,0.8]. This can be interpreted as evidence that the extracted data itself is of very rough quality. Again, this might be due to the quality of the dictionaries used to extract this data. Bad quality of the extracted data would explain why we observe such deteriorations, especially for the mixture model.

Interval	USFD-News	Cumulative	USFD-News
0.8 – 0.7	9	>0.7	9
0.7 – 0.6	30	>0.6	39
0.6 – 0.5	58	>0.5	97
0.5 – 0.4	112	>0.4	209
0.4 – 0.3	342	>0.3	551
0.3 – 0.2	2448	>0.2	2999
0.2 – 0.1	6471	>0.1	9470

Table 23. Statistics about experiments for Lithuanian->Romanian.

The experiments shown in Table 24 have to compete against the baseline which reached a BLEU score of 9.91. All but one system perform badly, losing at least 0.2 BLEU points to the baseline. Although we have one improved system using the interval of [0.6,0.5], there are only 58 additional sentences in this chunk. When we see this in context to the OOV counts from Section 4.2, so this is not a meaningful result.

Table 24 BLEU scores for experiments for Lithuanian->Romanian.

Interval	Interpolated LM – USFD-News	Cumulative	Interpolated LM – USFD-News
0.8 – 0.7	*9.13*	>0.7	*9.13*
0.7 – 0.6	*9.43*	>0.6	*9.32*
0.6 – 0.5	**10.11**	>0.5	*9.13*
0.5 – 0.4	*9.67*	>0.4	*9.02*
0.4 – 0.3	*9.55*	>0.3	***9.69***
0.3 – 0.2	*8.99*	>0.2	*9.33*
0.2 – 0.1	*9.54*	>0.1	*9.24*

In Figure 7 we can see that there is no clear tendency for the extracted data. Due to the small amount of data, it is not possible to say with certainty that using data with a particular threshold will improve translation quality.

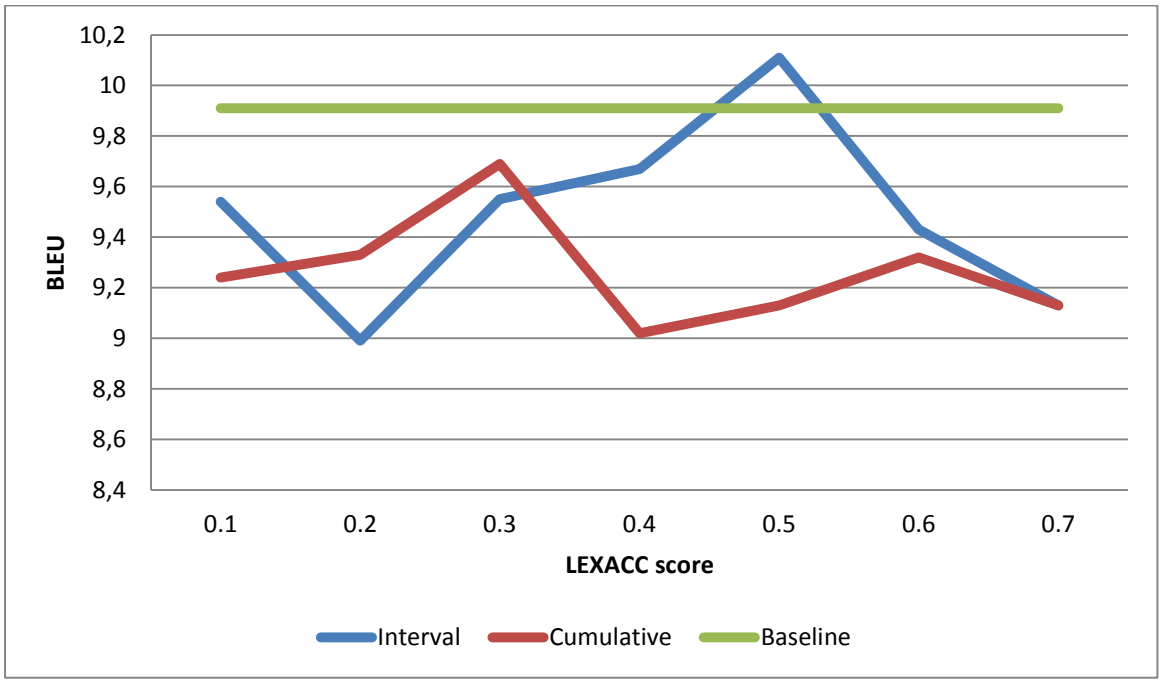

Figure 7 BLEU scores for experiments for Lithuanian->Romanian.

4.3.5. Romanian→German

Romanian→German is the last language pair we examined in our staggered experiments. Our working hypothesis is the degradations are caused by the same reasons as for Lithuanian→Romanian: the baseline corpus is also DGT/JRC, but it's also the smallest corpus based on DGT/JRC with only 615,336 lines. The extracted data from the USFD-News corpus comprises 10,227 lines (1.66%). As Table 25 shows, there are no extracted sentence pairs with a LEXACC score above 0.8.

Table 25 Statistics about experiments for Romanian->German.

Interval	USFD-News	Cumulative	USFD-News
0.8 – 0.7	8	>0.7	8
0.7 – 0.6	30	>0.6	38
0.6 – 0.5	55	>0.5	93
0.5 – 0.4	135	>0.4	228
0.4 – 0.3	804	>0.3	1032
0.3 – 0.2	3965	>0.2	4997
0.2 – 0.1	5230	>0.1	10227

Here, too, we see similar amounts for data as for Lithuanian→Romanian. The baseline BLEU score is 10.22, but when we compare this to the results of our experiments, we see that for Romanian→German the interpolated models fare much better. All but one improves over the baseline, although the cumulative results show a less clear tendency.

Table 26 BLEU scores for experiments for Romanian->German.

Interval	Interpolated LM – USFD-News	Cumulative	Interpolated LM – USFD-News
0.8 – 0.7	10.77	>0.7	**10.77**
0.7 – 0.6	10.54	>0.6	*9.72*
0.6 – 0.5	**11.21**	>0.5	10.67
0.5 – 0.4	*10.00*	>0.4	10.30
0.4 – 0.3	11.12	>0.3	*9.64*
0.3 – 0.2	10.57	>0.2	*9.90*
0.2 – 0.1	10.40	>0.1	*9.56*

Although we only add 55 sentences in the best-performing experiment, we observe an increase of 1 BLEU point. This is a significant improvement, but the small size of the data does not allow us to generalise the influence of this particular improvement. On the other hand, we can see the general tendency that BLEU increases with a higher LEXACC score. There's an outlier at [0.5,0.4], though, which indicates that this batch of data is of very bad quality. This may explain why the overall performance drops.

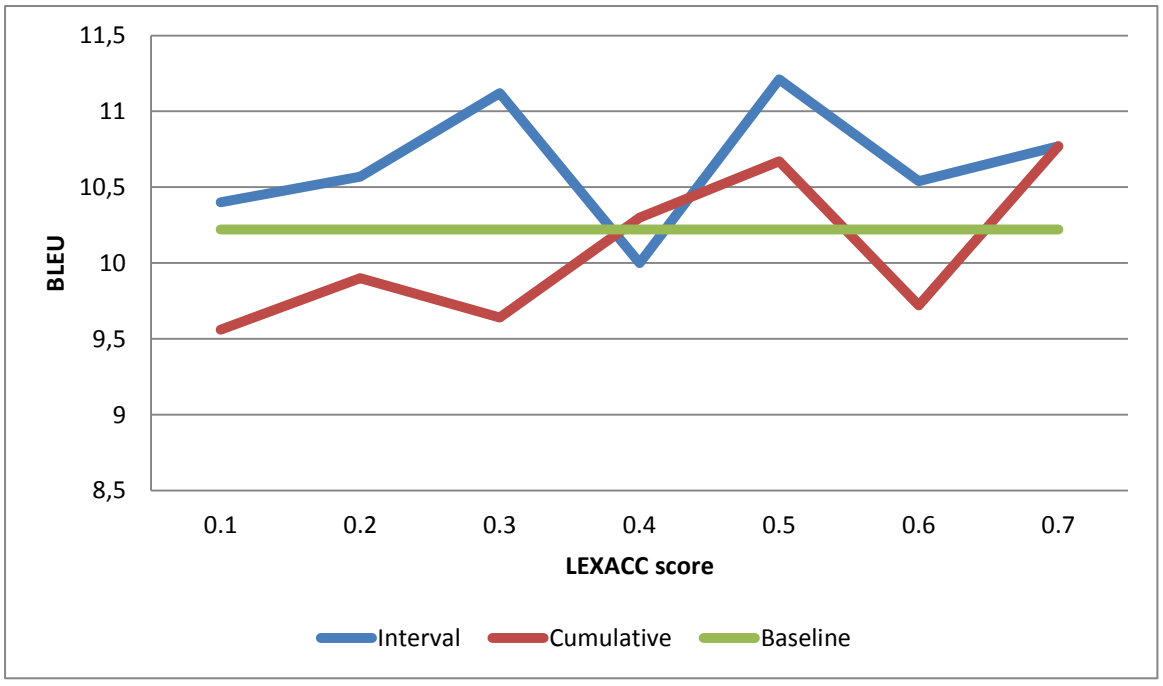

Figure 8 BLEU scores for experiments for Romanian->German.

5. Conclusions

In this deliverable we reported our experiments to enrich baseline SMT systems by using additional data extracted from comparable corpora. We wanted to investigate two questions:

1. Does the data extracted from comparable corpora help us to improve translation quality as measured by BLEU?
2. Does the LEXACC score correlate to the BLEU score?

Concerning the first question, this is the last step in evaluating the ACCURAT toolkit (see D2.6). The toolkit has already been evaluated at various points throughout the processing pipeline, such as checking the comparable corpora and also examining the data extracted by LEXACC. As the final goal of the toolkit is to provide data that helps to improve the BLEU score, we applied state-of-the-art approaches to make the best use of the new additional data, namely by a) interpolating language models while adding the additional data to the training corpus, and b) using mixture models.

For most of the language pairs we investigated, these approaches led to an improvement in BLEU score, as shown in Table 8. Although the amount of extracted data was very small compared to the training corpora, we manage to improve baseline systems based on DGT/JRC by .50 BLEU points on average. For systems using SETimes, the improvement is much higher, as here we add substantial amounts of data to the training corpus.

As for the second question, we see a slight correlation between LEXACC and BLEU for some language pairs. We do not take Lithuanian→Romanian and Romanian→German into consideration, as the amount of extracted data for these language pairs is too small to allow us to generalise over the translation results.

For English→Latvian and English→Romanian, we see that the best BLEU results are achieved when using the LEXACC threshold that achieves the best F1 measure. Although the results for English→Lithuanian do not match up in this manner, this is still a strong indication that this threshold is useful when filtering extracted data to receive the data that balances best comparability of the sentences and the amount of extracted data, as we have to work with the dichotomy that only using the sentences with a high confidence that they will be parallel to strongly comparable will result in a very small amount of data, but setting the confidence low to receive more data will give us more sentence pairs that are not parallel or strongly comparable.

The exact thresholds vary quite a lot depending on the language pair involved and even between the exact corpus used: for English→Lithuanian, the best threshold for the USFD-News data is 0.6, whereas the USFD-Wiki data performs best when we only use the data with a LEXACC score above 0.9. This makes it very difficult for any user to predict which threshold to choose—they will have to perform their own individual analysis of the extracted data to find the threshold.

In summary, the data extracted by LEXACC helps to improve the BLEU score for many language pairs. Although the user will have to find out the optimal threshold for their type of corpus, the LEXACC score can be used to filter the extracted data.

6. List of tables

7. Table of Figures