# Multi-source annotation projection of coreference chains: assessing strategies and testing opportunities

**Yulia Grishina** and **Manfred Stede**
Applied Computational Linguistics
FSP Cognitive Science
University of Potsdam
grishina|stede@uni-potsdam.de

## Abstract

In this paper, we examine the possibility of using annotation projection from multiple sources for automatically obtaining coreference annotations in the target language. We implement a multi-source annotation projection algorithm and apply it on an English-German-Russian parallel corpus in order to transfer coreference chains from two sources to the target side. Operating in two settings – a low-resource and a more linguistically-informed one – we show that automatic coreference transfer could benefit from combining information from multiple languages, and assess the quality of both the extraction and the linking of target coreference mentions.

## 1 Introduction

While monolingual coreference resolution systems are being constantly improved, multilingual coreference resolution has received much less attention in the NLP community. Most of the coreference systems can only work on English data and are not ready to be adapted to other languages. Developing a coreference resolution system for a new language from scratch is challenging due to its technical complexity and the variability of coreference phenomena in different languages, and it depends on high-quality language technologies (such as mention extraction, syntactic parsing, named entity recognition) as well as gold standard data, which are not available for a wide range of languages.

However, this can be alleviated by using cross-lingual projection which allows for transferring existing methods or resources across languages. There have been some influential work on annotation projection for different NLP tasks which performed quite well cross-lingually, e.g. for semantic role labelling (Akbik et al., 2015) or syntactic parsing (Lacroix et al., 2016). At the same time, several recent studies on annotation projection for coreference have proven it to be a more difficult task than POS tagging or syntactic parsing, which is hard to be tackled by projection algorithms. These works are limited to the existing multilingual resources (mostly newswire, mostly CoNLL 2012 (Pradhan et al., 2012)) and, surprisingly, are not even able to beat a threshold of 40.0 F1 for coreference resolvers trained on projections only. The best-performing system based on projection achieves 38.82 for English-Spanish and 37.23 for English-Portuguese F1-score (Martins, 2015), while state-of-the-art monolingual coreference systems are already able to achieve 64.21 F-score for English (Wiseman et al., 2016). While being quite powerful for other tasks, annotation projection is less successful for coreference resolution. Therefore, our question is, how can the quality of annotation projection be improved for the task of coreference resolution?

In our opinion, projection from multiple source languages can be a long-term solution, assuming that we have access to two or more reliable coreference resolvers on the source sides. Our idea is that multi-source annotation projection for coreference resolution would grant a bigger pool of potential mentions to choose from, which can be beneficial for overcoming language divergences. Therefore, the main goals of this study are: (a) to explore different strategies of multi-source projection of coreference chains on a small experimental corpus, and (b) to evaluate the projection errors and assess the prospects of this approach for multilingual coreference resolution.

This paper is structured as follows: The related work is discussed in Section 2, and the dataset is presented in Section 3. The methodology adapted



*Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), co-located with EACL 2017*, pages 41–50,
Valencia, Spain, April 4, 2017. 

for our experiments is explained in Section 4. We then analyse the projection errors and evaluate the target annotations (Section 5). Finally, Section 6 summarises the outcomes of this study, and Section 7 concludes.

## 2 Related work

Annotation projection is a method of automatically transferring linguistic annotations from one language to the other in a parallel corpus. It was first applied in the pilot work of Yarowski et al. (2001) who used this technique to induce POS and Named Entity taggers, NP chunkers and morphological analyzers for different languages. In particular, they used labelled English data and an aligned parallel corpus to automatically create mappings between the annotations from the source side and the corresponding aligned words on the target side, and exploited the resulting annotations to train their systems.

Thereafter, projection has been widely used as a method in cross-lingual NLP, and several studies on annotation projection targeted cross-lingual coreference resolution. In particular, automatic annotation transfer was first applied to coreference chains by Postolache et al. (2006) who used a projection method and filtering heuristics to support the creation of a coreference corpus in a new language. The evaluation of projected annotations against a small manually annotated corpus exhibited promising 63.88 and 82.6 MUC and B-cubed scores respectively. Subsequently, Souza and Orăsan (2011) went one step further and made an attempt to project automatically produced annotations, and used projected data to train a new coreference resolver, which, however, resulted in a poor coreference resolution quality due to low-quality annotations on the source side.

The next steps in projecting coreference included several translation-based approaches. The difference is that the target text is first translated into the source language, on which coreference resolution is performed; after that, the source coreference chains can be projected back to the target side. This approach was used, for example, by Rahman and Ng (2012) to train coreference resolvers for Spanish and Italian using English as the source language, achieving an average F1 of 37.6 and 21.4 for Spanish and Italian respectively in a low-resource scenario, and much better scores of 46.8 and 54.9 F1 using only a mention extractor. Similarly, Ogrodniczuk (2013) experimented with translation-based projection for English and Polish using only a mention extractor. The evaluation of the quality of the projected annotations on manually annotated data showed 70.31 F1.

The most recent application of projection to coreference is due to Martins (2015) who experimented with transferring automatically produced coreference chains from English to Spanish and Portuguese, and subsequently trained target coreference resolvers on the projected data, combining projection with posterior regularization. His approach shows competitive results in a low-resource setting, with the average of 38.82 F1 for coreference resolution systems trained on projections for Spanish and 37.23 for Portuguese, as compared to the performance of fully supervised systems: 43.93 and 39.83 respectively.

The idea of using multiple sources for annotation projection was also initially considered by Yarowsky et al. (2001) who used multiple translations of the same text to improve the performance of the projected annotations for several NLP tasks. Furthermore, multi-source projection has been extensively explored for multilingual syntactic parsing. The best unsupervised dependency parsers nowadays rely on annotation projection (Rasooli and Collins, 2015; Johannsen et al., 2016). To our knowledge, there has been no attempt to apply multi-source annotation projection to the task of coreference resolution so far.

## 3 Data

For our experiments, we have chosen a trilingual parallel annotated coreference corpus of English, German and Russian from (Grishina and Stede, 2015). This corpus was annotated with coreference chains according to the guidelines described in (Grishina and Stede, 2016) which are largely compatible to the coreference annotations of the OntoNotes corpus (Pradhan and Xue, 2009). The corpus is annotated with full coreference chains, excluding singletons[1]. The major differences to OntoNotes are: (a) annotation of NPs only, but not of verbs that are coreferent with NPs, (b) inclusion of appositions into the markable span and not marking them as a separate relation, (c) marking relative pronouns as separate markables, and (d)

[1]Mentions of the entities that appear in the text only once.

| | News | | | Stories | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | EN | DE | RU | EN | DE | RU | EN | DE | RU |
| Sentences | 229 | 229 | 229 | 184 | 184 | 184 | 413 | 413 | 413 |
| Tokens | 6033 | 6158 | 5785 | 2711 | 2595 | 2307 | 8744 | 8753 | 8092 |
| Markables | 560 | 586 | 604 | 466 | 491 | 471 | 1026 | 1077 | 1075 |
| Chains | 115 | 133 | 133 | 40 | 40 | 45 | 155 | 173 | 178 |

Table 1: Corpus statistics for English, German and Russian

annotation of pronominal adverbs[2] in German if they co-refer with an NP.

Since the corpus was already aligned bilingually for two language pairs – English-German and English-Russian – we first align the German-Russian corpus at the sentence level using LF Aligner[3] and then select parallel sentences present in all the three languages. This method reduces the average number of sentences per language by 5% and the average number of coreference chains per language by 6% (as compared to the corpus statistics published by Grishina and Stede (2015)). Then we re-run GIZA++ word aligner (Och and Ney, 2003) on the resulting sentences for all the language combinations with German and Russian as targets.

The statistics of the experiment corpus after selecting only trilingual sentences are presented in Table 1.

## 4 Experiments

Combining information coming from two or more languages is a more challenging task as compared to single-source projection where one just transfers all the information from one language to the other. For coreference, this task is non-trivial (as opposed to, for instance, multi-source projection of POS information where an intuitive majority voting strategy could be chosen), since we cannot operate on the token level and even not on the mention level: We cannot implement a strategy to choose e.g. the most frequent label for a token or a sequence of tokens (coreferent/non-coreferent), since they belong to mention clusters which are not aligned on the source sides. In other words, if mention $x_a$ belongs to chain $A$ in the first source language and mention $y_b$ belongs to chain $B$ in the second source language, and they are projected onto the same mention $z_{ab}$ on the target side, we do not know whether both target chains $A'$ and $B'$ projected from $A$ and $B$ respectively and both containing the mention in question are equal or not, as we cannot rely on chain IDs which are not common across languages. Therefore, we have to operate on the chain level and first compare projected coreference chains. We treat coreference chains as clusters, measure the similarity between them and use this information to choose between them or combine them together in the projection.

Projecting coreference chains (=clusters of mentions) from more than one language, we can have the following cases:

(a) Two chains are identical (contain all the same mentions);

(b) Two chains are disjoint (contain no same mentions);

(c) Two chains overlap (contain some identical mentions).

While cases (a) and (b) are quite straightforward, case (c) is more difficult since we have to determine whether to treat these chains as being equal or not.

Following the work of (Rasooli and Collins, 2015), we rely upon two strategies – concatenation and voting – to process coreference chains coming from two sources. Since we only have two sources, instead of voting we implement intersection. In the case of coreference, we can enrich annotations from one language with the annotations from the other one or create a completely new set out of two projection sets. In particular, we experiment with several naive methods and evaluate their quality, and then combine them with each other in order to find the optimal strategy.

We implement the following methods:

(1) **Concatenation:** Data is obtained from each of the languages separately and then concatenated.

   (a) `add`: Disjoint chains present in only one language are added to the projected

[2] Adverbs that are formed by combining a pronoun and a preposition, e.g. *therefor*.

[3] https://sourceforge.net/projects/aligner/

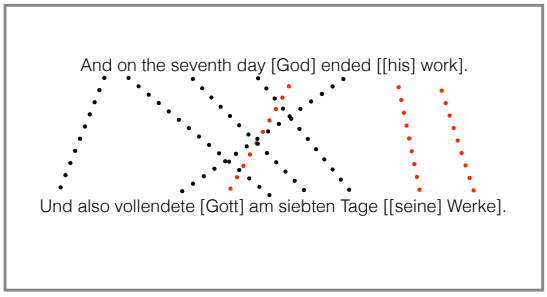



Figure 1: Direct projection algorithm

chains from the other language. Typically, we would take projected annotations for the best-scored language and enrich them with annotations from the less-scored language.

(b) `unify-concatenate (u-con)`: Overlapping chains from both languages are merged together: If chain $A$ and chain $B$ overlap, we concatenate the mentions from both chains that form a new chain $AB$.

(2) **Intersection:** Projected annotations are obtained by intersecting projections coming from two sources.

(a) `intersect (int)`: The intersection of coreference chains present in both languages is chosen[4].

(b) `unify-intersect (u-int)`: The intersection of the mentions for overlapping chains is chosen: If chain $A$ and chain $B$ overlap, we intersect the mentions from both chains that form a new chain $AB$.

We use the following formula to estimate the overlap between two coreference chains:

$$\frac{2|A \cap B|}{|A| + |B|}, \quad (1)$$

where $A$ and $B$ are the number of mentions for coreference chains in question. We experiment with different values of overlap and choose the best one for each of the methods[5]. For `u-int`, we perform intersection of mentions for all the chains with mention overlap over 0.05. For `u-con`, we select chains with 0.5 overlap value for German and 0.7 for Russian. If the overlap is less than these values, we treat these chains as disjoint.

Each of the methods is applied in the following settings:

1. Setting 1: no additional linguistic information available. In this setting, we use only word alignments to transfer information from one language to the other.

2. Setting 2: a mention extractor is available. Relying on the output of the MATE dependency parser[6] (Bohnet, 2010) for German and the MALT dependency parser[7] (Nivre et al., ) for Russian[8], we automatically extract all mentions that have nouns, pronouns or pronominal adverbs as their heads. Thereafter, we map the output of the projection algorithm to the extracted mentions. We modify the mapping strategy described in (Rahman and Ng, 2012), mapping (a) projected markables that are identical to the extracted mentions, (b) projected markables that share the same right boundary with the extracted mentions, (c) markables that are spanned by the extracted mentions, (d) all other markables for which no corresponding mentions were found. Once a markable is mapped to a mention, we discard this mention, to ensure that it is not mapped to any other markable. For Russian, we skip step (b), which leads to better scores.

As the baseline, we select a single-source projection method. We re-implement a simple direct projection algorithm as described in (Postolache et al., 2006) and (Grishina and Stede, 2015), and we run it for the English-German, English-Russian, German-Russian and Russian-German language pairs, since we are not interested in projecting into English. The direct projection is illustrated in Fig.1 where coreference mentions *God*, *his* and *his work* are transferred to the German side via word alignments. Then, we run the algorithm in the two settings described above. Note that the projection results for setting 1 are slightly lower as compared to the results reported in (Grishina and Stede, 2015): we did not rely on intersective word

[4]Imagining we have more than two source languages, we could implement a more sophisticated voting scheme

[5]We use part of the corpus to determine optimal thresholds and the other one to obtain the results.

[6]https://code.google.com/archive/p/mate-tools/

[7]http://www.maltparser.org

[8]Using the model provided by Sharoff and Nivre (2011)

| | MUC | | | $B^3$ | | | $CEAF_m$ | | | Avg. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| EN→DE | 57.6 | 46.0 | 51.1 | 47.3 | 35.6 | 40.4 | 61.1 | 49.7 | 54.7 | 55.3 | 43.8 | 48.7 |
| RU→DE | 43.3 | 28.4 | 34.1 | 33.3 | 18.9 | 23.5 | 46.2 | 32.7 | 38.1 | 40.9 | 26.7 | 31.9 |
| EN,RU→DE: | | | | | | | | | | | | |
| - add | 52.7 | 46.1 | 49.1 | 41.5 | 36.5 | 38.6 | 53.5 | 51.2 | 52.2 | 49.2 | 44.6 | 46.6 |
| - int | 46.7 | 2.5 | 4.5 | 82.3 | 3.1 | 5.6 | 87.5 | 3.6 | 6.5 | **72.2** | 3.1 | 5.5 |
| - u-con | 56.0 | 48.8 | 52.1 | 44.5 | 38.8 | 41.3 | 59.4 | 51.9 | 55.3 | 53.3 | **46.5** | **49.6** |
| - u-int | 64.7 | 26.1 | 36.7 | 58.6 | 18.7 | 27.3 | 65.7 | 32.4 | 43.1 | 63.0 | 25.7 | 35.7 |
| EN→DE+*ment* | 66.7 | 53.1 | 59.0 | 54.8 | 41.6 | 47.0 | 68.1 | 55.3 | 61.1 | 63.2 | 50.0 | 55.7 |
| RU→DE+*ment*: | 43.6 | 28.5 | 34.2 | 34.3 | 19.1 | 24.0 | 47.1 | 33.4 | 38.8 | 41.7 | 27.0 | 32.3 |
| EN,RU→DE | | | | | | | | | | | | |
| - add+*ment* | 60.0 | 53.1 | 56.2 | 47.0 | 42.7 | 44.3 | 57.9 | 56.9 | 57.2 | 55.0 | 50.9 | 52.6 |
| - int+*ment* | 56.7 | 3.6 | 6.3 | 96.7 | 4.5 | 7.9 | 97.8 | 4.9 | 8.6 | **83.7** | 4.3 | 7.6 |
| - u-con+*ment* | 66.1 | 55.7 | 60.4 | 53.4 | 45.0 | 48.6 | 67.4 | 57.4 | 61.9 | 62.3 | **52.7** | **57.0** |
| - u-int+*ment* | 73.7 | 29.6 | 41.7 | 68.1 | 21.6 | 31.3 | 73.6 | 36.1 | 48.0 | 71.8 | 29.1 | 40.3 |

Table 2: Results for German

| | MUC | | | $B^3$ | | | $CEAF_m$ | | | Avg. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| EN→RU | 71.3 | 55.1 | 62.0 | 61.2 | 43.0 | 50.3 | 71.5 | 56.6 | 63.1 | 68.0 | 51.6 | 58.5 |
| DE →RU: | 59.1 | 32.0 | 41.3 | 46.8 | 19.6 | 27.3 | 57.3 | 35.1 | 43.3 | 54.4 | 28.9 | 37.3 |
| EN,DE→RU | | | | | | | | | | | | |
| - add | 67.8 | 55.5 | 60.9 | 55.8 | 43.7 | 48.8 | 64.8 | 57.9 | 61.0 | 62.8 | **52.4** | 56.9 |
| - int | 87.5 | 3.0 | 5.9 | 85.0 | 4.3 | 8.2 | 85.0 | 4.8 | 9.0 | **85.8** | 4.0 | 7.7 |
| - u-con | 70.6 | 55.7 | 62.2 | 60.1 | 43.6 | 50.4 | 71.0 | 57.1 | 63.2 | 67.2 | 52.2 | **58.6** |
| - u-int | 81.6 | 29.3 | 42.9 | 74.8 | 19.5 | 30.6 | 77.6 | 35.5 | 48.6 | 78.0 | 28.1 | 40.7 |
| EN→RU+*ment* | 71.6 | 55.4 | 62.3 | 61.7 | 43.2 | 50.6 | 72.0 | 57.1 | 63.5 | 68.4 | 52.4 | 58.8 |
| DE →RU+*ment* | 59.2 | 32.0 | 41.4 | 47.5 | 19.7 | 27.6 | 57.9 | 35.4 | 43.8 | 54.9 | 29.0 | 37.6 |
| EN,DE→RU | | | | | | | | | | | | |
| - add+*ment* | 68.0 | 55.7 | 61.1 | 56.7 | 44.1 | 49.3 | 65.1 | 58.3 | 61.4 | 63.3 | **52.7** | 57.3 |
| - int+*ment* | 87.5 | 2.4 | 4.7 | 85.0 | 3.5 | 6.6 | 85.0 | 3.9 | 7.5 | **85.8** | 3.3 | 6.3 |
| - u-con+*ment* | 70.9 | 56.0 | 62.4 | 60.9 | 43.9 | 50.8 | 71.5 | 57.5 | 63.6 | 67.7 | 52.5 | **59.0** |
| - u-int+*ment* | 82.2 | 29.2 | 42.9 | 76.4 | 19.4 | 30.6 | 78.7 | 35.7 | 49.0 | 79.1 | 28.1 | 40.8 |

Table 3: Results for Russian

alignments, since we were not interested in maximizing Precision at the cost of low Recall. Our goal was to obtain balanced scores to base our experiments upon.

The results for the baselines and the experiments are presented in Table 2 and Table 3. We compute the standard coreference metrics using the latest version of the CoNLL-2012 official scorer[9]. We also compute the average scores for all the coreference metrics.

## 5 Error analysis

We perform the error analysis by evaluating the projection quality for each of the methods described above. We first look at the common and distinct chains projected from two languages, and thereafter we evaluate the projection quality for different NP types and for the mentions of different length.

**Common chains projected from two sources (`int`).**

To analyse the common chains projected from two sources into German and Russian, we extract these chains from the target annotations and discard the singletons (if any). We compute the average chain length – 2.75 and 2.13 for German and Russian respectively – and look at the types of mentions that occur in these chains. Interestingly, string match is the most frequent type, e.g. 'Indien' - 'Indien', 'Афганистане' - 'которого' - 'Афганистане' ('Afghanistan' - 'which' - 'Afghanistan'). Named Entities form 46% of all the markables, followed by pronouns, which are 27% of all markables. Still, the Recall numbers are too low (3.1 and 4.0 for German and Russian) to apply this method on a small corpus.

[9]https://github.com/conll/reference-coreference-scorers

**Distinct chains added from one source to the other (`add`).** We examine the chains added from the less-scored language to the best-scored one by extracting these chains separately and computing their Precision. The results for both languages exhibit low Precision: 20.0 Precision for mention extraction and 15.0 average Precision for coreference, and 14.0 and 7.0 for German and Russian respectively. These numbers are too low to improve the projection performance in a low-resource setting.

**Evaluation by NP type (`u-int, u-con`).** In order to evaluate the projection quality for different NP types, we computed the distribution of types for the source and target annotations. For that reason, we POS-tagged the corpus using TreeTagger[10] (Schmid, 1995) with the pre-trained models for German and Russian. Subsequently, we extract the gold and the projected markables and compare them according to their types.

For German, we distinguish between the most frequent markable types: common NPs, Named Entities, personal, possessive, demonstrative and relative pronouns. For Russian, we only distinguish between the common NPs, Named Entities and pronouns, relying on the tagset available for TreeTagger[11]. Table 4 shows the distribution of all markables, regardless of whether they are correct or incorrect, for both the `u-int, u-con` settings. We do not show the percentage for the markables that are not of the types described below, but count them in the total numbers.

Interestingly, the percentage of NPs + Named Entites (computed together) and pronouns for both projections and for both methods is quite comparable (59.0 vs. 59.3, 54.7 vs. 58.4). However, the percentage of common NPs and Named Entities in German and Russian (computed separately) is not the same, the reason being different POS tagsets for the two languages used by TreeTagger. For Russian, a large amount of proper names were identified as common nouns, e.g. 'India', 'Mumbai', 'Hamas' etc. For German, these were identified as Named Entities.

Based on these observations, we compute the projection accuracy of each NP type as the number of correct markables of this type divided by the total number of projected markables of the same type. Table 5 shows the projection accuracy for both settings. According to these results, in the knowledge-lean approach, NPs are the less reliable projected type for German as compared to Named Entities, which is due to the fact that most of them lose their determiners at the alignment stage. For Russian, both NPs and Named Entities show similar results of over 80% with the `u-int` method. With the `u-con` method, all the scores are a bit lower due to lower Precision obtained by concatenation. As one can see from columns 3 and 4, it is possible to significantly improve the NP identification accuracy for German by using only a mention extractor: over 17% for both methods. However, this is not the case for Russian, where NP extraction relying on word alignment does not produce that much noise: the improvement is around 0.5-2.8%.

Pronouns exhibit the best projection accuracy for both languages. For German, the highest scores are achieved by the projection of possessive (97.1), personal (95.1) and relative (81.8) pronouns. Demonstrative pronouns show the lowest score (50.0) due to their scarcity in the gold and projected data. In setting 2, we can only achieve little improvement for different pronoun types, except for personal pronouns for German that exhibit lower accuracy.

These results explain the better projection quality when projecting to Russian as compared to projecting to German, since all the projected types show fair projection accuracy. Conversely, German NPs show poorer accuracy, while constituting almost one third of all the projected markables, which inevitably leads to lower Precision and Recall scores.

**Evaluation by mention length (`u-int`).** Finally, we compare mentions according to the number of tokens they consist of. Fig. 2a and Fig. 2b show the overall amount of tokens and the number of correct tokens of this length for German and Russian respectively in the `u-int` setting, in which higher Precision results were achieved. For German, the number of correct mentions gradually decreases up to the length of 5; after that, only one or no correct mentions are to be found in the target annotations. For Russian, the situation is almost the same, except for the mentions with length of 3, which are mostly incorrect.

[10] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[11] http://corpus.leeds.ac.uk/mocky/

| | unify-int | | | | u-con | | | |
|---|---|---|---|---|---|---|---|---|
| | →DE # | →DE % | →RU # | →RU % | →DE # | →DE % | →RU # | →RU % |
| NPs | 146 | 29.6 | 286 | 58.8 | 264 | 28.4 | 450 | 52.2 |
| Named Entities | 145 | 29.4 | 26 | 0.05 | 245 | 26.3 | 53 | 6.2 |
| Pronouns | | | 113 | 23.3 | | | 237 | 27.5 |
| -Personal pronouns | 82 | 16.6 | - | - | 143 | 15.4 | - | - |
| -Possessive pronouns | 35 | 7.1 | - | - | 69 | 7.4 | - | - |
| -Demonstrative pronouns | 2 | 0.4 | - | - | 5 | 0.5 | - | - |
| - Relative pronouns | 11 | 2.2 | - | - | 12 | 1.3 | - | - |
| Total | 494 | 100 | 486 | 100 | 931 | 100 | 862 | 100 |

Table 4: Distribution of all projected markables by type for u-int and u-con methods

| | u-int | | u-con | | u-int+*ment* | | u-con+*ment* | |
|---|---|---|---|---|---|---|---|---|
| | →DE % | →RU % | →DE % | →RU % | →DE % | →RU % | →DE % | →RU % |
| NPs | 53.4 | 82.5 | 53.0 | 77.8 | 72.0 | 85.3 | 70.1 | 78.3 |
| Named Entities | 91.0 | 92.3 | 82.0 | 88.7 | 95.2 | 92.3 | 84.1 | 88.7 |
| Pronouns | | 92.0 | | 89.9 | | 92.9 | | 90.3 |
| Personal pronouns | 95.1 | - | 95.1 | - | 87.8 | - | 92.3 | - |
| Possessive pronouns | 97.1 | - | 94.2 | - | 97.2 | - | 98.6 | - |
| Demonstrative pronouns | 50.0 | - | 40.0 | - | 100.0 | - | 40.0 | - |
| Relative pronouns | 81.8 | - | 83.3 | - | 100.0 | - | 100.0 | - |

Table 5: Projection accuracy for u-int and u-con methods

## 6 Discussion

Analysing the results for multi-source projection for both target languages, one can see that the scores achieved are quite comparable: the highest Precision of 83.7/85.8 for German/Russian and the highest Recall of 52.7 for both. Looking at the u-int method in setting 2, we still see that Precision is somewhat higher for Russian than for German (79.1 vs. 71.8 respectively). Overall, the best F1-scores for both languages are 57.0/59.0 German/Russian.

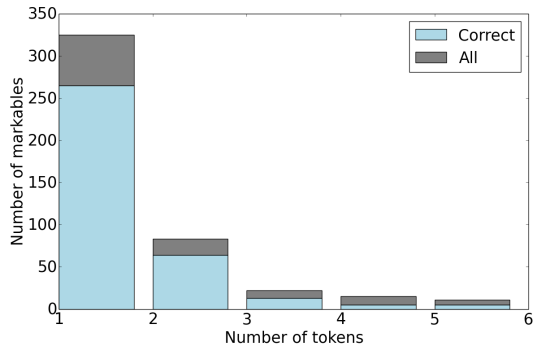
Importantly, for both target languages and in both settings, the multi-source projection results outperform the single-source results in terms of Precision or Recall; however, still not both simultaneously. In particular, the u-con method exhibits higher F1 scores as compared to single-source projection (55.0 vs. 57.0 for German and 58.8 vs. 59.0 for Russian).

As for the different projection methods, the results show that the balance between Precision and Recall scores is quite stable in both settings. In particular, concatenating mentions in overlapping chains (u-con) resulted in the most balanced Precision and Recall scores for both German and Russian. Furthermore, Precision can be improved in two ways: by taking the intersection of chains coming from two languages and by taking the intersection of mentions in the overlapping chains in two languages. While the first scenario is more unrealistic, leading to extremely low Recall numbers, the second scenario returns much better results in terms of both Precision and Recall.

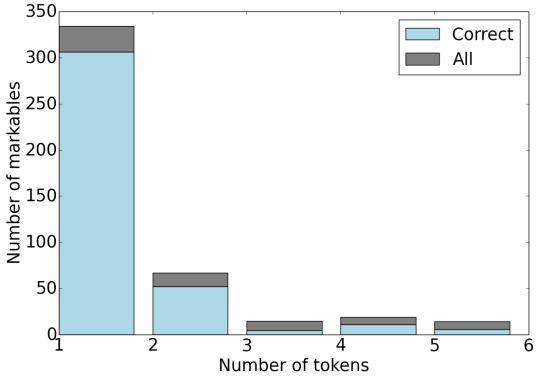
Comparing our results to the most closely related work of Grishina and Stede (2015), we can see a large improvement in the projection quality for English-German in terms of both Precision and Recall already in the knowledge-lean setting: best Precision of 72.2 vs. 78.4/53.4 news/stories[12] respectively, and best Recall of 46.5 vs. 41.4/45.9. In setting 2, the results are even better: 83.7 and 57.0. As for Russian, we conclude that the multi-source approach leads to a slight improvement of projection results in terms of Precision (best Precision of 85.8 for settings 1,2 vs. 73.9/84.6), but not in terms of Recall (52.4 for setting 1 and 52.7 for setting 2 vs. 58.3/59.0), which is also due to the fact that the single-source projection performed slightly worse in the absence of intersective alignments.

Interestingly, the results for single-source projection also show that different directions of projection are not equally good: Projection from English still shows the best results, while Projection from German to Russian and from Russian to German exhibit much lower F1 numbers. In our opinion, the fact that projection results with language other than English as source are much lower had a negative impact on the multi-source projection, since adding lower-quality annotations

[12] Mind that stories constitute 30% of the corpus, therefore we consider our overall results higher.




(a) German




(b) Russian

Figure 2: Overall number of mentions and the number of correct mentions according to the number of tokens

leads to a decrease in both Precision and Recall scores. Therefore, concatenation of the two projections with one of them being of lower quality results in a slight drop in Precision and does not improve the Recall numbers significantly. Using projections of similar quality and more languages would result in better overall scores.

Automatic mention extraction and the mapping of target mentions to the extracted mentions to a high degree supported the identification of mentions and hence coreference scores for the English-German language pair. For Russian, conversely, this method only helped to a small extent, the reason being already high Precision scores achieved by projecting through word alignment. The qualitative analysis has shown that incorrectly identified mentions were of wrong part-of-speech (e.g. verbs, therefore it was not possible to map them to the automatically extracted mentions) or were no markables in the gold annotations.

In sum, our results have shown that projecting from two sources rather than one helps both to improve Precision and Recall. However, improving Precision appears to be an easier task than improving Recall. Achieving higher Recall seems to be a more difficult and expensive task as compared to eliminating noisy alignments and ensuring correct mention boundaries. If a potential target mention is absent on the source sides, it can hardly be recovered in the resulting annotations.

## 7 Conclusions

In this work, we examined the multi-source approach to projecting coreference annotations in a low-resource and a more linguistically-informed setting by implementing a direct projection algorithm and several methods for combining annotations coming from two sources. Comparing our results to a single-source approach, we observed that the former is able to outperform the latter one, both in terms of Precision and Recall. Specifically, our results suggest that the concatenation of coreference chains coming from two sources exhibits the highest balanced Precision and Recall scores, while the intersection helps to achieve the highest Precision.

We further analyzed the errors both quantitatively and qualitatively, focusing on the nature of the projected chains coming from both languages and the projection accuracy of different coreference mention types. Our results showed that noun phrases are more challenging for the projection algorithm than pronouns, and, as a by-product, we found that using automatic mention extraction to a large extent supports the recovery of target markables expressed by common noun phrases for German. However, this is not necessarily the case for Russian, for which using higher quality word alignments is more effective.

Having tested and assessed several methods of two-source annotation projection, we envision our future work on automatic annotation transfer in combining annotations coming from more than two source languages. Furthermore, we are interested in adapting a similar approach for projecting automatic annotations, which, in our opinion, could support the creation of a large-scale coreference corpus, suitable for the training of coreference resolvers in new languages.

## References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu

Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China, July. Association for Computational Linguistics.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd international conference on computational linguistics*, pages 89–97. Association for Computational Linguistics.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 14–22, Beijing, China, July. Association for Computational Linguistics.

Yulia Grishina and Manfred Stede, 2016. *Parallel coreference annotation guidelines*. Unpublished Manuscript[13].

Anders Johannsen, Željko Agić, and Anders Søgaard. 2016. Joint part-of-speech and dependency projection from multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–566, Berlin, Germany, August. Association for Computational Linguistics.

Ophélie Lacroix, Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Frustratingly easy cross-lingual transfer for transition-based dependency parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1058–1063, San Diego, California, June. Association for Computational Linguistics.

André F. T. Martins. 2015. Transferring coreference resolvers with posterior regularization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1427–1437, Beijing, China, July. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of 5th international conference on Language Resources and Evaluation (LREC)*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

[13]Available at https://github.com/yuliagrishina/CORBON-2017-Shared-Task/blob/master/Parallel_annotation_guidelines.pdf

Maciej Ogrodniczuk. 2013. Translation-and projection-based unsupervised coreference resolution for Polish. In *Language Processing and Intelligent Information Systems*, pages 125–130. Springer.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *Proceedings of 5th international conference on Language Resources and Evaluation (LREC)*.

Sameer S. Pradhan and Nianwen Xue. 2009. Ontonotes: The 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12, Boulder, Colorado, May. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, July. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Translation-based projection for multilingual coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338. Association for Computational Linguistics.

Helmut Schmid. 1995. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the ACL SIGDAT-Workshop*. Association for Computational Linguistics.

Serge Sharoff and Joakim Nivre. 2011. The proper place of men and machines in language technology: Processing russian without any linguistic knowledge. In *Proc. Dialogue 2011, Russian Conference on Computational Linguistics*.

José Guilherme Camargo Souza and Constantin Orăsan. 2011. Can projected chains in parallel corpora help coreference resolution? In *Discourse Anaphora and Anaphor Resolution Colloquium*, pages 59–69. Springer.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San

Diego, California, June. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.