Improving Verb Phrase Extraction from Historical Text by use of Verb Valency Frames

Eva Pettersson
Department of Linguistics and Philology, Uppsala University
eva.pettersson@lingfil.uu.se

Abstract

In this paper we explore the idea of using verb valency information to improve verb phrase extraction from historical text. As a case study, we perform experiments on Early Modern Swedish data, but the approach could easily be transferred to other languages and/or time periods as well. We show that by using verb valency information in a post-processing step to the verb phrase extraction system, it is possible to remove unprobable complements extracted by the parser and insert probable complements not extracted by the parser, leading to an increase in both precision and recall for the extracted complements.

1 Introduction

In the Gender and Work project (GaW), historians are building a database with information on what men and women did for a living in the Early Modern Swedish society, i.e. approximately 1550–1800 (Ågren et al., 2011). This information is currently extracted by researchers manually going through large volumes of text from this time period, searching for relevant text passages describing working activities. In this process, it has been noticed that working activities often are described in the form of verb phrases, such as hugga ved ("chop wood"), sälja fisk ("sell fish") or tjäna som piga ("serve as a maid"). Based on this observation, Pettersson et al. (2012) developed a method for automatically extracting verb phrases from historical documents by use of spelling normalisation succeeded by tagging and parsing. Using this approach, it is possible to correctly identify a large proportion of the verbs in Early Modern Swedish text. Due to issues such as differences in word order and significantly longer sentences than in present-day Swedish texts (combined with sentence segmentation problems due to inconsistent use of punctuation), it is however still hard for the parser to extract the correct complements associated with each verb.

In this work we propose a method for improving verb phrase extraction results by providing verb valency information to the extraction process. We describe the effect of removing unprobable complements from the extracted verb phrases, as well as adding probable complements based on verb valency information combined with words and phrases occurring in close context to the head verb. In Section 2, we present related work, whereas our approach to improving verb phrase extraction for historical text is presented in Section 3. The data used in our experiments are described in Section 4, and the evaluation method in Section 5. Finally, the model selection process is described in Section 6, the results are presented in Section 7, and conclusions are drawn in Section 8.

2 Related Work

Baldwin and Villavicencio (2002) explored three methods for automatically extracting English verb-particle constructions from the WSJ section of the Penn Treebank: 1) based on part-of-speech tagging output, 2) based on chunking output, and 3) by use of a chunk grammar. The first two methods resulted in high precision but rather low recall, whereas the third method yielded high recall at the cost of a
relatively low precision. By combining these three methods, a precision of 85.9% and a recall of 87.1% were achieved.

Baldwin (2005) tried a combination of linguistic and statistical tests for automatic extraction of prepositional verbs. For a gold standard of 135 prepositional verbs derived from the BNC corpus, an f-score of 45% was reported.

Jakubčík and Kovář (2013) introduced a verb valency-based method for improving Czech parsing. Their experiments are based on the Synt parser, which is a head-driven chart parser with a handcrafted meta-grammar for Czech, producing a list of ranked phrase-structure trees as output. They used two different dictionaries with valency information to rerank the suggested parses in accordance with the valency frames suggested for the verb in the dictionaries. Evaluation was performed on the Brno Phrasal Treebank using the leaf-ancestor assessment metric, and an improvement from 86.4% to 87.7% was reported for the highest-ranked tree when comparing the Synt parser in its original setting to the inclusion of valency frames for reranking of the output parses.

Pettersson et al. (2013) presented an approach to automatic annotation of historical text, including verb phrase extraction. In this approach, the historical spelling is translated to a modern spelling employing character-based statistical machine translation (SMT) techniques, before tagging and parsing is performed by use of standard natural language processing (NLP) tools developed for present-day language. Applying this approach to Early Modern Swedish, the number of correctly identified verb complements (including partial matches) increased from 32.9% for the text in its original spelling to 46.2% for the text in its automatically modernised spelling. Earlier work by the same authors showed that using contemporary valency dictionaries to remove extracted complements not adhering to the valency frame of a specific verb had a positive effect on verb phrase extraction precision (Pettersson et al., 2012).

3 Approach

In this work, we adopt the verb phrase extraction method presented in Pettersson et al. (2013), where verbs and complements are extracted from historical text based on output from tagging and parsing tools developed for present-day Swedish. In addition to their approach, we also include a post-processing step, removing and/or inserting verbal complements based on the valency frame of the head verb.

The full process is illustrated in Figure 1, where the first step is tokenisation of the historical source text by use of standard tools. The tokenised text is then to be linguistically annotated in the form of tagging and parsing. To the best of our knowledge, there is however no tagger nor parser available that has been trained on Early Modern Swedish text. Since these tools are sensitive to spelling, the tokenised text is therefore normalised to a more modern spelling by use of character-based SMT methods, before tagging and parsing is performed using tools trained for modern Swedish. For tagging, we use HunPOS (Halácsy et al., 2007) with a Swedish model based on the Stockholm-Umeå corpus, SUC (Ejerhed and Källgren, 1997). For parsing, we use MaltParser version 1.7.2 (Nivre et al., 2006a) with a pre-trained model based on the Talbanken section of the Swedish Treebank (Nivre et al., 2006b).

After tagging and parsing, the annotations given by the tagger and the parser are projected back to the text in its original spelling, resulting in a tagged and parsed version of the historical text, from which the verbs and their complements are extracted. The complements included for extraction are the following: subject (for passive verbs only, where the subject normally corresponds to the direct object in an active verb construction), direct object, indirect object, prepositional complement, infinitive complement, subject predicative, verb particle, and reflexive pronoun. As mentioned, we also add a post-processing filter as a complementary step, using valency information to modify the complements suggested by the parser.

3.1 Deletion of Unprobable Complements

As discussed in Section 1, certain characteristics of historical text make it difficult for the parser to correctly extract the complements of a verb. Therefore, we add valency information in a post-processing step, filtering away extracted complements that do not conform to the valency frame of the verb. A similar idea was presented in Pettersson et al. (2012),
Figure 1: Method overview.

where filtering was based on valency frames given in two contemporary dictionaries, i.e. Lexin\(^1\) and Parole\(^2\). However, some word forms in historical text are not frequent enough in contemporary language to occur in modern dictionaries. Examples from the GaW training corpus are *absentera* (old word for "be absent"), *umgälla* (old word for "suffer for"), and *ärna* (old word for "intend to"). Moreover, the meaning of verbs tend to change over time, and it is not obvious that verb valency frames for present-day Swedish also holds for historical Swedish. An example from the GaW corpus is the verb *slå* ("hit") which in both Lexin and Parole is listed as a monolingual verb ("to hit someone"). In the GaW corpus however, it is repeatedly used as a ditransitive verb, as in *Sedhan hadhe Erich OluffSon slagit Pelle Pederssson tre blånder* ("Then Erich Oluff-Son had hit Pelle Pederssson three bruises"). In our approach to deletion of unprobable complements, we therefore base the valency frames not only on the contemporary Lexin and Parole valency dictionaries, but also on the verbal complements occurring in the training part of the GaW corpus.

Deletion experiments are performed for all complement types extracted from the parser except for subjects, since a verb is typically expected to have a subject. We present deletion experiments for the following five settings:

1. **Lexin**
   For each extracted complement, if the head verb is present in the Lexin valency dictionary and the valency frame in Lexin does not allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase.

2. **Parole**
   For each extracted complement, if the head verb is present in the Parole valency dictionary and the valency frame in Parole does not allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase.

3. **GaW Corpus**
   For each extracted complement, if the head verb is present in the training part of the GaW corpus, and none of the occurrences in the corpus contain a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase.

4. **All combined**
   For each extracted complement, if the head verb is present in all three resources mentioned:

\(^1\)http://spraakbanken.gu.se/lexin/valens_lexikon.html
\(^2\)http://spraakbanken.gu.se/swe/resurs/parole
above, and none of these resources allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase. Likewise, if the head verb is present in only two of these three resources, and none of these two resources allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase. Finally, if the head verb is present in one resource exclusively, and this resource does not allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase.

5. All one-by-one
For each extracted complement, if the head verb is present in the best-performing resource, i.e. the resource yielding the highest complement extraction f-score in the first three experiments, and the valency frame in this resource does not allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase. Otherwise, if the head verb is present in the second best-performing resource, and the valency frame in this resource does not allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase. Only if the head verb is not present in any of the two best-performing resources, the third resource is consulted.

3.2 Insertion of Probable Complements
Apart from filtering away unlikely complements extracted by the parser, we also aim at inserting probable complements not found by the parser, by searching the parsed sentence for words and phrases that match the valency frame of the head verb, but which have not been extracted by the parser. Since the word order is more varying in Early Modern Swedish than in present-day Swedish, all complements are searched for both to the left and to the right of the head verb.

In the insertion experiments, we focus on phrasal verbs in the broader sense, including particles, reflexives, and prepositional complements. We believe that these complement types are relatively easy to recognise in a sentence. Furthermore, if for example a reflexive pronoun is found within a reasonable distance from the head verb, and the valency frame suggests a reflexive pronoun, then the probability that this reflexive belongs to the verb is rather high. The same argument holds for prepositional phrases containing the expected preposition to form a prepositional complement, and for prepositions or adverbials identical to a particle expected by the valency frame of the head verb. For direct and indirect objects on the other hand, even if we find a noun phrase close to the verb, it would still be hard to determine whether this noun phrase actually corresponds to a direct or indirect object, since noun phrases may occur with many different functions in a clause, and the word order is not fixed, especially not for historical text. Therefore we would run a high risk of extracting for example the subject noun phrase instead of the direct or indirect object noun phrase. Furthermore, direct objects are not always expressed in the form of noun phrases, but are quite often expressed as for instance clauses, as in the following example from the GaW corpus: *fordra at Barnet skal döpas hemma* (“demand that the Child should be christened at home”). Similarly, subject predicatives may also be expressed in varying ways and infinitive complements are often ambiguous to other functions. Thus, these categories are excluded from the deletion experiments.

In accordance with the arguments given above, the following three experiments are performed for insertion of probable complements:

1. Insertion of prepositional complement
If the valency frame of the head verb (in any of the three valency resources) allows for a prepositional complement, and a prepositional phrase containing the expected preposition is found within a reasonable distance either to the left or to the right of the head verb, this prepositional phrase is added to the extracted verb phrase with a prepositional complement label.

2. Insertion of particle
If the valency frame of the head verb (in any of the three valency resources) allows for a particle, and a word that is identical to the expected particle and tagged as preposition or adverb is found within a reasonable distance either to the
left or to the right of the head verb, this preposition or adverb is added to the extracted verb phrase with a particle label.

3. Insertion of reflexive

If the valency frame of the head verb (in any of the three valency resources) allows for a reflexive pronoun, and the word form sig ("oneself", with an alternative historical spelling sigh) is found within a reasonable distance either to the left or to the right of the head verb, this word form is added to the extracted verb phrase with a reflexive label.

4 Data

Verb valency frames are extracted from three sources: the contemporary Lexin valency dictionary, the contemporary Parole valency dictionary, and the training and development parts of the GaW corpus of Early Modern Swedish court records and church documents. Evaluation is performed on the evaluation part of the GaW corpus. All the verbs in the GaW corpus have been manually annotated as such, and all complements adhering to the verbs have been annotated with labels denoting subject (for passive verbs only), direct object, indirect object, prepositional complement, infinitive complement, subject predicative, verb particle, and reflexive pronoun. Furthermore, the training and development parts of the corpus have been annotated with information on the manually modernised spelling for each original word form occurring in the text.

Both in the Lexin dictionary and in the Parole dictionary, verb valency frames are connected to the present tense form of the verb only, without information on other inflectional forms of the verb. In the verb phrase extraction process however, we need to connect whatever inflectional form of the verb that is used in the sentence to the correct valency frame. For broader coverage of the valency dictionaries, the present tense forms were therefore expanded to other inflectional forms based on the Saldo dictionary and the SUC corpus. The Saldo dictionary is a dictionary of present-day Swedish word forms, with morphological and inflectional information (Borin et al., 2008). By comparing the present tense verb form in Lexin or Parole to the Saldo dictionary, it is thus possible to extract a lemma corresponding to the verb form, and from that lemma all the inflectional forms adhering to that lemma. For verb forms not found in the Saldo dictionary, the SUC corpus was consulted. Since this corpus has been manually annotated with lemma information, all inflectional forms of the same lemma occurring in the corpus may thus be extracted. For Lexin and Parole verb forms not found in neither Saldo nor SUC, only the present tense form of the verb is stored with its corresponding valency frame.

For the GaW corpus, we have a similar problem in that only those verb forms that occur in the corpus will be assigned a valency frame, and if several forms of the same verb occur in the corpus, these will be assigned valency frames separate from each other. To deal with this, we use the same method of comparison to Saldo and SUC for retrieving the full set of word forms associated with a verb form, assigning the same valency frame to all verb forms belonging to the same lemma. In this process, we use the manually normalised form of each verb for comparison towards Saldo and SUC, to avoid mismatches due to spelling variation in the historical corpus.

Table 1 shows the number of verb forms found in Saldo and SUC respectively, during the process of expanding the valency frames to more inflectional forms. The GaW corpus has been divided into training (train), development (dev) and test sets, where the training part is the same data set as was used for training and tuning in Pettersson et al. (2013), and the development set is the same data set as was used for evaluation in the same paper. Since the test set will only be used for evaluation, no expansion to inflectional forms is needed for this particular data set.

<table>
<thead>
<tr>
<th></th>
<th>Verbs</th>
<th>Saldo</th>
<th>SUC</th>
<th>Not found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexin</td>
<td>3,281</td>
<td>3,181</td>
<td>33</td>
<td>67</td>
</tr>
<tr>
<td>Parole</td>
<td>4,304</td>
<td>4,263</td>
<td>26</td>
<td>15</td>
</tr>
<tr>
<td>GaW Train</td>
<td>1,329</td>
<td>1,168</td>
<td>14</td>
<td>147</td>
</tr>
<tr>
<td>GaW Dev</td>
<td>1,410</td>
<td>1,245</td>
<td>15</td>
<td>150</td>
</tr>
<tr>
<td>GaW Test</td>
<td>987</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 1: Verb forms found in Saldo and SUC during the process of expanding the valency frames to more inflectional forms.
Table 2 lists the total number of entries in the language resources, before and after word form expansion. We will use the training part of our corpus as a basis for valency frames during model selection, where the development part is used for repeated testing. In the final evaluation, the training and development sets are merged to a combined valency resource, whereas evaluation scores are given for the test part of the corpus.

<table>
<thead>
<tr>
<th></th>
<th>verb forms</th>
<th>expanded forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexin</td>
<td>3,281</td>
<td>42,545</td>
</tr>
<tr>
<td>Parole</td>
<td>4,304</td>
<td>32,640</td>
</tr>
<tr>
<td>GaW Train</td>
<td>1,329</td>
<td>10,032</td>
</tr>
<tr>
<td>GaW Dev</td>
<td>1,410</td>
<td>10,394</td>
</tr>
<tr>
<td>GaW Test</td>
<td>987</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 2: Number of verb forms in the language resources, before and after word form expansion.

5 Evaluation

Evaluation is performed in terms of precision, recall and f-score based on the extracted complements, where the baseline case is the original verb phrase extraction system without any of the above specified amendments. We define true positives as correctly extracted complements. Likewise, false positives are complements extracted by the system that are not present in the gold standard, whereas false negatives are complements that are present in the gold standard but not extracted by the system. Since we are specifically aiming at extracting the correct complements, intransitive verbs that were also identified as intransitive by the extraction system will not contribute to the set of true positives. Intransitive verbs for which the system has extracted complements will however contribute to the set of false positives, whereas verbs identified as intransitive by the system though complements are present in the gold standard will add to the set of false negatives.

We also make a distinction between labelled and unlabelled precision and recall, where labelled precision and recall requires that the correct label for the complement has been assigned, i.e. direct object, prepositional complement etc, whereas unlabelled precision and recall only concerns the extracted word sequences, regardless of what label the parser has assigned to the complement.

Since the overall aim of the verb phrase extraction process is to present to historians text passages that may be of interest, partial matches are also regarded as true positives, as these would still point the user to the right text passage. True positives thus include the following cases, with authentic examples from the GaW corpus:

- **Exact match**
  - Gold complement: *2 klimpar smör*  
  - Extracted complement: *2 klimpar smör*  
  - “2 lumps of butter”

- **Substring type A**
  - Gold complement: *de penningar och medel*  
  - Extracted complement: *medel*  
  - ”(the money and) resources”

- **Substring type B**
  - Gold complement: *detta*  
  - Extracted complement: *detta after honom*  
  - ”this (after him)”

- **Overlap**
  - Gold complement: *försvägat ock förtrygt*  
  - Extracted complement: *nogh försvägat*  
  - ”(probably) weakened (and oppressed)”

6 Model Selection

In the model selection phase, we try different strategies for deletion and insertion of complements, using the training part of the corpus as a basis for valency frames, and the development part of the corpus for testing. As a baseline we use the verb phrase extraction system without the additional deletion or insertion of complements.

6.1 Deletion of Unprobable Complements

For deletion of unprobable complements, we first need to decide which of the five settings listed in Section 3.1 that should be chosen. We therefore ran experiments where deletion is performed for all complement types (except subject), evaluating the results for each setting separately. The results for unlabelled complement extraction are summarized in Table 3.
<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>53.30</td>
<td>51.22</td>
<td>52.24</td>
</tr>
<tr>
<td>Lexin</td>
<td>59.76</td>
<td>35.44</td>
<td>44.49</td>
</tr>
<tr>
<td>Parole</td>
<td>55.22</td>
<td>38.49</td>
<td>45.36</td>
</tr>
<tr>
<td>GaW corpus</td>
<td>57.51</td>
<td>46.77</td>
<td>51.59</td>
</tr>
<tr>
<td>All combined</td>
<td>56.62</td>
<td>47.76</td>
<td>51.81</td>
</tr>
<tr>
<td>All one-by-one</td>
<td>57.64</td>
<td>46.01</td>
<td>51.17</td>
</tr>
</tbody>
</table>

Table 3: Unlabelled results for deletion of unprobable complements with different settings.

As seen from the results, all settings improve precision as compared to the baseline system. However, recall varies to a great extent. For the largest resource, i.e. the Lexin dictionary, precision is the highest, but recall is very low. This indicates that a great amount of verb forms are found in the Lexin dictionary, but with valency frames that do not correspond to the way the verbs are used in historical texts, meaning that complements are erroneously deleted. This confirms our initial hypothesis that due to language change, contemporary dictionaries are not sufficient for guiding a parser with valency information. Further arguments for this hypothesis is the fact that even though the GaW training corpus is by far the smallest valency resource, using only this resource for defining verb valency frames results in a substantially higher f-score value than using Lexin or Parole. In fact, the f-score results for using the GaW corpus only are almost as high as for using all resources combined.

Since all methods improve precision as compared to the baseline, we choose the combined method for further experiments, since this method has the highest recall and also the highest f-score.

In the next round of experiments, we want to find out which complement types should be candidates for deletion. The hypothesis is that some complement types may be more thoroughly covered in the valency resources than others. If so, deletion of complements may only be a successful method for some complement types, whereas others should be left unmodified in the deletion process. To test this hypothesis, we tried deletion for each complement type separately, keeping only those that improve f-score as compared to the baseline system. These experiments were run with the combined setting, in accordance with the arguments given above, and the results are presented in Table 4, where it can be noticed that only deletion of direct objects and subject predicatives are successful in improving the f-score value as compared to the baseline. Keeping these two categories as candidates for deletion, a precision of 54.96% is achieved, with a recall of 50.25%, as compared to the baseline precision of 53.30 and recall of 51.22.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>53.30</td>
<td>51.22</td>
<td>52.24</td>
</tr>
<tr>
<td>A) direct object</td>
<td>54.29</td>
<td>50.62</td>
<td>52.39</td>
</tr>
<tr>
<td>B) indirect object</td>
<td>53.37</td>
<td>50.92</td>
<td>52.12</td>
</tr>
<tr>
<td>C) prep compl</td>
<td>56.06</td>
<td>47.51</td>
<td>51.43</td>
</tr>
<tr>
<td>D) inf compl</td>
<td>53.34</td>
<td>51.02</td>
<td>52.15</td>
</tr>
<tr>
<td>E) subject pred</td>
<td>53.93</td>
<td>50.84</td>
<td>52.34</td>
</tr>
<tr>
<td>F) particle</td>
<td>53.45</td>
<td>50.84</td>
<td>52.11</td>
</tr>
<tr>
<td>G) reflexive</td>
<td>53.19</td>
<td>50.50</td>
<td>51.81</td>
</tr>
<tr>
<td>A + E</td>
<td>54.96</td>
<td>50.25</td>
<td>52.50</td>
</tr>
</tbody>
</table>

Table 4: Unlabelled results for deletion of unprobable complements for the setting "all combined", varying the complements included for deletion.

### 6.2 Insertion of Probable Complements

As described in Section 3.2, the insertion experiments are targeted at particles, reflexives, and prepositional objects. Whenever the valency frame of the head verb in the extracted phrase allows for a complement of the specified type, the parsed sentence is searched for words and phrases matching the complement at hand. In the insertion experiments, we tried the following enhancements of the original insertion strategy:

1. Inclusion of stopwords, for which no complements are to be added. The set of stopwords were empirically defined as word forms belonging to any of the lemmas `vara` ("be"), `bli` ("become"), `ha` ("have") and `finnas` ("exist").
2. Restriction on what tokens are allowed to occur between the head verb and the candidate complement in order for the complement to be inserted, so that punctuations are not permitted.
3. Inclusion of a distance threshold, defining how many tokens that may come in between the
head verb and the candidate complement, in order for the complement to be inserted. We tried a number of different thresholds, out of which a threshold of 5 tokens turned out to yield the best results.

The insertion results are presented in Table 5, showing that without any restrictions in the insertion process, recall can be increased from 51.22% to 53.63%. This is however at the expense of a substantial drop in precision from 53.30% to 37.47% as compared to the baseline system. Restrictions in the form of A) stopwords for which no complements are inserted, B) prohibition of punctuation between the head verb and the candidate complement, and C) defining a threshold for how many tokens are allowed to occur between the head verb and the candidate complement, all had a positive effect on precision and f-score. Thus, in the best setting, i.e. where all three restrictions are implemented, a precision of 52.57% is achieved, with a recall of 52.45%.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>53.30</td>
<td>51.22</td>
<td>52.24</td>
</tr>
<tr>
<td>Original</td>
<td>37.47</td>
<td>53.63</td>
<td>44.12</td>
</tr>
<tr>
<td>A) Stopwords</td>
<td>45.69</td>
<td>53.41</td>
<td>49.25</td>
</tr>
<tr>
<td>B) Punctuation</td>
<td>47.81</td>
<td>53.04</td>
<td>50.29</td>
</tr>
<tr>
<td>C) Threshold</td>
<td>51.42</td>
<td>52.54</td>
<td>51.97</td>
</tr>
<tr>
<td>A + B + C</td>
<td>52.57</td>
<td>52.45</td>
<td>52.51</td>
</tr>
</tbody>
</table>

Table 5: Unlabelled results for insertion of probable complements.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>53.25</td>
<td>38.34</td>
<td>44.58</td>
</tr>
<tr>
<td>Deletion</td>
<td>54.75</td>
<td>37.97</td>
<td>44.84</td>
</tr>
<tr>
<td>Insertion</td>
<td>53.67</td>
<td>40.45</td>
<td>46.13</td>
</tr>
<tr>
<td>Delete + Insert</td>
<td>55.12</td>
<td>40.08</td>
<td>46.41</td>
</tr>
</tbody>
</table>

Table 6: Labelled and unlabelled results for complement extraction.

decreasing precision, demonstrating that inserting complements introduces true positives to a higher extent than false positives, which is satisfactory. The best precision is achieved when both deletion and insertion of complements are performed, yielding a precision of 63.04%, as compared to 61.82% for the baseline system. This setting also yields an improvement in both precision and recall, as compared to the baseline.

8 Conclusion

We have presented a method for improving verb phrase extraction from historical text, by automatically deleting unprobable verbal complements extracted by the parser, while at the same time inserting probable complements not extracted by the parser. Our approach is based on verb valency frames rendered from historical corpora and from contemporary valency dictionaries, where the historical corpus had the largest positive effect even though the contemporary dictionaries covered more verb forms. This indicates that since language changes over time, valency frames for present-day language may not be enough to cover the syntax in historical text. By automatically deleting and inserting complements based on a combination of the historical corpus and the contemporary dictionaries, an increase in both precision and recall is achieved, as compared to the baseline system.

For historians working with old texts, there is a need for NLP tools to effectively search large
volumes of text automatically for text passages of special interest. We believe our method for verb phrase extraction from historical text to be a useful tool for this purpose, a hypothesis that has also been confirmed by the historians themselves. Still, there is room for improvement, since the best precision achieved for complement extraction is 63.04%, with a recall of 49.74%. In the current approach, verb valencies are exploited in a post-processing phase, with the original extracted verb phrases as input. Future work includes to explore the possibility of providing valency information already in the parser training phase, enriching the part-of-speech tags with information on whether a certain verb is likely to occur with for example a particle or prepositional complement. The hypothesis is that a parser trained on this kind of data will be keen to search harder for the expected complements.

References


