THE NEED OF MANUAL ANNOTATIONS IN AN AUTOMATED PROCESS

BY SOPHIE VAN DUIN AND KIM VAN DER MEER

VU VRIJE UNIVERSITEIT AMSTERDAM
Faculteit der Letteren

INTRODUCTION

This research is about the manual annotations of a Humanities project: SERPENS. When Van Erp et al. (2018) conducted their research 3 questions arose, the aim of this research is to answer these 3 questions.

SERPENS is a project about historical ecology. The project consists of several elements, Van Erp et al (2018) created a workflow. In the part of manual annotations it is important to get the perfect balance between annotation effort and performance. During the project of Van Erp et al. (2018) the categories of annotating have changed. First there were 10 categories and right now there are 3. This could be less time consuming, yet will there still be a good balance? Not only species that have been annotated before play a role, also new species play a role in this research. Another aspect is that the annotations can be done on paragraph or document level. When annotating on document level it would be less time consuming. This led to 3 research questions:

To what extent are effort and performance in balance with each other?

Should there be annotated on document or paragraph level?

What effect will annotating a new specie have on the categories? Should these categories be changed?

DATA & METHOD

The data used is the KB newspaper archive, this a collection of Dutch historical newspapers online. The period of the newspapers investigated is 1800 till 1940 because those have good OCR quality and there was a high level of biological interest this years.

For the annotations there were consisting guidelines and a program. The annotations are made in a Post Gres database with a web front end. The annotation classes are: No animal (OCR mistake), Animal and Figurative. Two different species are annotated: the polecat and the sturgeon. For the polecat there are expert annotations available, yet the sturgeon has not been annotated before. In this research there are two non-expert annotators and this will give the possibility to compare the results in multiple ways. For the polecat the annotations can be compared with each other or with the expert annotations and for the sturgeon the agreement between the two non-expert annotators can be calculated. This comparison will be done with Cohen's Kappa which will give the Inter-Annotator Agreement (IAA) as outcome.

RESULTS

Tables 1 and 2 show the inter-annotator agreement that was calculated with Cohen's Kappa.

	Annotator 1	Annotator 2
Matching annotations	167 of 183	168 of 184
Cohen's Kappa	0,24	0,35

Table 1 IAA of both annotators with the expert

As table 1 shows, both of the annotators did not score very high on the Cohen's Kappa with the expert. Even though the annotations were done very similarly, the agreement was between 0.21 and 0.40 in both cases which is classified as fair agreement.

	Polecat	Sturgeon
Matching annotations	310 of 324	200 of 218
Cohen's Kappa	0,29	0,78

Table 2 IAA between annotator 1 and 2

Table 2 shows that for the polecat it is the same case as the agreement with the expert, the agreement is fair while there are a lot of matching annotations. However, for the sturgeon the agreement is substantial.

The annotating task was not very complicated, especially the polecat annotations did not take a very long time. The sturgeon was a bit more difficult to annotate. Together the annotators marked 1200 snippets. This took the annotators 20 hours, more or less. This comes down to 1 annotation per minute.

Both the polecat and the sturgeon were annotated in the same three categories. Classifying the snippets within these categories was suitable for both species. However, the category 'animal' grew much bigger than any of the other categories. For the polecat, there was an average of 309,5 out of 324 annotations as 'animal' and for sturgeon this average was 163,5 out of 218.

CONCLUSION/DISCUSSION

As Cohen's Kappa showed, there was only a fair agreement between the annotators and the expert annotations. This while there was a very high number of matching annotations. This can be caused by the following aspects. Firstly, the expert annotations were done within ten categories while the other annotators only used three categories. Secondly, Cohen's Kappa relies on coincidence. This means that, by coincidence, every category has as much chance to be chosen as any other. This does not apply in this case, since the category 'animal' was chosen much more than the other categories. For the agreements between the two non-experts, for the polecat it is the same case as described above. As for the sturgeon, the agreement is substantially high which is caused by a better distribution within the categories.

As for the time consumption, the annotating task did not take much time given the fact that the annotators were not experts on the data. So there was a good balance between effort and performance. The reduction from 10 to 3 categories definitely made the process easier. Perhaps the process could be more accelerated if the annotators are trained more.

In words of the level of annotations, both annotators concluded this should not be changed. Annotating on paragraph level is clear and makes it accessible to assign a snippet to a category. Annotating on document level would cause overgeneralizations.

As noticed when annotating and also when reading the results, the category 'animal' is by far the largest category. However, within this categorie there is no difference between the animal as pest or nuisance specie and the animal in a different appearance such as a product. When distinguishing between these two appearances, the SERPENS project will benefit from this. This way the results from the annotations give much more useful information.

DISCUSSION

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities