

# Exploration of Event Extraction Techniques in Late Medieval and Early Modern Administrative Records

Ismail Prada Ziegler

*Digital Humanities, University of Bern, Switzerland*

*Department of History, University of Basel, Switzerland*

## Abstract

While an increasing amount of studies exploring named entity recognition in historical corpora are published, application of other information extraction tasks such as event extraction remains scarce. This study explores two accessible methods to facilitate the detection of events and the classification of entities into roles: rule-based systems and RNN-based machine learning techniques. We focus on a German-language corpus from the 15th-17th c. and property purchases as the event types. We show that these relatively simple methods can retrieve useful information and discuss ideas to further enhance the results.

## Keywords

information extraction, historical data, digital history, machine learning

## 1. Introduction

Among historical documents from the late medieval and early modern periods administrative records are one of the most prevalent types of source material. These documents exhibit a high density of information and often display some degree of standardisation within collections. These traits make them ideal candidates for digital methods of information extraction and analysis.

However, applying digital information extraction techniques to historical documents presents numerous challenges. Annotated historical datasets are limited both in size and in number, and variations in grammar and spelling due to the lack of standardisation pose significant obstacles. Despite these difficulties, notable advancements have been made in the field due to growing interest in digital history and digital humanities. An overview of recent studies concerning named entity recognition can be found in [3].


This paper contributes to this evolving field by presenting a case study on extracting event information from historical land registers. In our project *Economies of Space* we work to digitize these registers and explore the potential of extracting information such as entities, relations and events.<sup>1</sup> Our goal is to create a knowledge base where the individual histories of persons, properties, and organizations can be explored, as well as to enable distant reading methods of analysis. In [6] we demonstrated that robust named entity recognition is possible for our

---

*CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark*

✉ ismail.prada@unibe.ch (I. Prada Ziegler)

ORCID 0000-0003-4229-8688 (I. Prada Ziegler)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

<sup>1</sup><https://dg.philhist.unibas.ch/en/bereiche/mittelalter/forschung/oekonomien-des-raums/>

data. In this study, we explore the potential of event extraction as a first step to investigate interactions between the found entities. We compare two methods: rule-based extraction and RNN-based machine learning. While this case study focuses on a narrow example, we hope that the findings of these experiments will benefit other teams working with similar datasets.

## 2. Dataset

### 2.1. The Historical Land Registers

The experiments were conducted with the Historical Land Registers of Basel.<sup>2</sup> This archival collection aimed to bring together excerpts from all archival documents which mention a property inside the old city of Basel. The content is a mix of legal and bookkeeping information relating to property ownership, rents, and transactions. Our project focuses on 80,000 excerpts from between 1400 and 1700, written in Early New High German.<sup>3</sup> Almost all excerpts are kept to a single sentence, even when describing complex events. For the remainder of this paper, the term "sample" will refer to an individual document within this collection. The documents were automatically transcribed with an average CER of 3.6%.

### 2.2. Entity Annotation

640 samples were annotated following the BeNASch guidelines.<sup>4</sup> BeNASch applies a nested entity representation, which means for each entity mention, a mention span (e.g., 'the house at the river', 'Hans Stuber, the tailor') is annotated as well as a *head* element (e.g., 'house', 'Hans Stuber'). All entity mentions that fall into one of the categories PER (persons), ORG (organizations), LOC (locations), or GPE (Geo-political entities), including pronouns, are annotated.

### 2.3. Event Annotation

The 640 samples also feature event annotation. We define an event as a "specific occurrence involving participants" following the ACE guidelines.<sup>5</sup> Only events that belong to categories which were determined in our project to be of interest to historical research are annotated. An event is characterized by two main elements: the trigger and the roles. The trigger represents a word or phrase around which the event is centered. Roles match entity annotations and describe the entities part in the event. See Appendix A for an annotation example.

---

<sup>2</sup><https://dls.staatsarchiv.bs.ch/records/1016781>

<sup>3</sup>Although as is always the case with copied documents, we must suspect that at least in some cases modifications and to some degree modernization of text took place. To answer the question "to what degree?" is part of our research project.

<sup>4</sup><https://dhbern.github.io/BeNASch/>

<sup>5</sup><https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf>. Similar guidelines have since been adopted in BeNASch.

## 3. Methodology

### 3.1. Data and Evaluation

**Table 1**

Number of occurrences of each role and trigger phrases.

Category	Count
seller	200
buyer	220
property	184
price	129
trigger	173

For the purposes of this study, we focus on the event type *property purchase*. It appears in 167 out of the 640 samples, making it comparatively frequent. We define the following roles for the event *property purchase*: seller (PER or ORG), buyer (PER or ORG), property (LOC), price (MONEY). Every role may appear multiple times, and only the property role must appear at least once. The total occurrences of each role are shown in Table 1. We implement 5-fold stratified cross-validation because our dataset is still extremely small, especially for machine-learning purposes. We split each fold 60/20/20% for training, validation and testing respectively. The results represent the average across the five folds. This dataset still contains all other event-annotated samples, but triggers and roles in those have been removed (we do this to evaluate if our systems can distinguish property purchase events from other events as well).

### 3.2. Rule-based System

#### 3.2.1. Trigger Detection

For each fold, we create a gazetteer of potential trigger phrases by counting the trigger-annotated phrases in that folds training-set. We exclude phrases which appear fewer than  $k$  times from the gazetteer. We then compare this gazetteer to the input samples and apply fuzzy-matching, using the *thefuzz* python library<sup>6</sup>, to mark one or multiple tokens as triggers. We allow our algorithm to detect multiple trigger phrases in a single sample. For each fold, we determine the minimum ratio for the fuzzy matching as well as the minimum frequency  $k$  by running different parameter combinations against the validation set and choosing the best result. The best parameters were either setting  $k$  to 3 and minimum fuzz-ratio to 0.8 or setting both parameters to 1.

We avoid some frequent problems with additional rules: 1. To prevent misidentification of the word "Kauf" in documents titled "Kauf-Urkunde" (purchase deed), we forbid the first token in a document to match a trigger. 2. Triggers may only match words outside of entity mentions, this prevents for example "verkauft" in "das verkauft Haus" (the sold house) from being identified as a trigger. 3. If one trigger follows another trigger without any entity mention in

---

<sup>6</sup><https://github.com/seatgeek/thefuzz>

between, we remove the second one (e.g. "es verkauft und gibt zu kaufen"). While these kinds of errors don't have a negative impact on the document classification or role detection, they distort the trigger detection scores to look more negative than they actually are. 4. In some cases two predicted triggers are separated by a person or organization mention. We reclassify the second trigger as a *helper* in that case. They are helpful information in the role detection and their use will be explained in the next section. 5. Finally, we remove triggers where a MONEY or TIME annotation is found between the trigger and a LOC. These cases indicate rent purchase documents which are very similar in language and structure to property purchase documents.

### 3.2.2. Role Detection

To detect roles, we apply a simple template system whenever a trigger is present. For property purchase documents, we identify three different kinds of structures (ignoring non-entity-mentions and non-values):

1. <SELLER><TRIGGER><BUYER><PROPERTY><MONEY>
2. <TRIGGER><SELLER><BUYER><PROPERTY><MONEY>
3. <BUYER><PROPERTY><TRIGGER><MONEY>

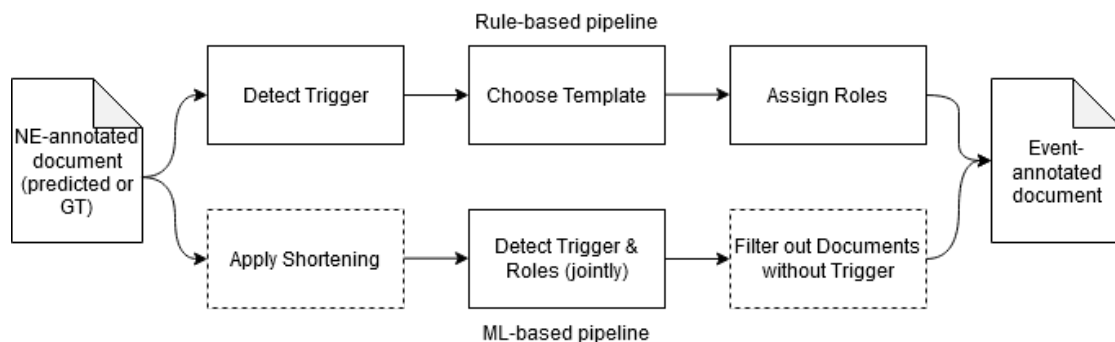
Templates 1 and 2 are usually found when the sale is the central event of the excerpt, while template 3 usually follows a seizure event, giving information who bought the property after it was seized and auctioned off. Sometimes roles are missing from the text, so we only require a trigger and at least one LOC-mention to apply a template. The template used is decided by looking at the differences: A LOC before the trigger implies template 3, otherwise check if a PER/ORG is present before the TRIGGER, if yes then template 1 is used, otherwise template 2. We can match mentions to roles due to the restrictions in their categories, as long as their position relative to the other roles and the trigger is correct: PER and ORG can only be SELLER and BUYER, while LOC can only be PROPERTY and MONEY can only be PRICE. One challenge is the distinction between SELLER and BUYER in template 2. To show what can already be done by simple means in this case study, we solve this by putting the first half of all PER/ORG mentions as SELLER and the second half of all PER/ORG mentions as BUYER (in case of an odd number of candidates, SELLER gets the additional one). If a helper-trigger is present, we use it to distinguish buyer and seller.

## 3.3. Machine-Learning System

### 3.3.1. Architecture

Our approach to event extraction by machine learning is inspired by previous successes to extract entities in pre-modern German texts [4][6]. Like entities, we can model roles and trigger as annotation spans in the text and apply a sequence tagging strategy (this is one common way to model event extraction [5]).

We implement our experiment using the FlairNLP framework [1]. For the language model, we stack a forward and backward model of contextual character embeddings [2] which we obtained by finetuning the de-model on all handwritten documents in the Historical Land



**Figure 1:** Overview of both strategies to recognize events. Dashed frames imply variants. Using the event recognition on plain texts is also possible, but prevents the rule-based approach and the shortening step.

Registers, including later than 1700 (appr. 9.14M token). Character-based embeddings have demonstrated robustness against the inherent variability of pre-modern German spelling and vocabulary [4]. For the event extraction, we train a sequence tagging model with the default settings of Flair (single-layered Bi-LSTM + CRF decoder).

### 3.3.2. Pretagging

To insert the information from the named entity annotation into the model, we add a prefix and suffix token to each entity mention. "Hans sold his house ." becomes "[B-PER] Hans [E-PER] sold [B-LOC] his house [E-LOC] ." For experiments focused only on role detection, we incorporate trigger information in the same manner. ("[B-SALE] sold [E-SALE]").

We conduct experiments with manually annotated tags as well as automatically predicted ones. The predicted annotations are trained as a Flair SequenceTagger as well, using the same language model as the event recognition. A separate model is trained for each fold so no data contamination occurs.

When pretagging is applied, the role detection is not required to match the whole span of the pretagged entity, instead it is trained to classify the prefix token (e.g. "[B-PER]") correctly. The training data is adjusted accordingly (see Appendix B for an example).

### 3.3.3. Variants

**Shortening:** To shorten our samples and possibly remove noise, we remove all tokens inside entity annotations which are not part of the *head*. This reduces our sample length by an average of about a third of all tokens. The NER models for this variant of pretagging were trained as described in [6].

**Document Filtering:** Because the system in initial tests often annotated roles in documents even when they (correctly) identified no trigger, we added a rule to disregard all role annotations in documents where no trigger is present.

Note that these variants do only apply to the machine learning strategy. Shortening is irrelevant to the rule-based strategy because trigger detection only happens outside of entity annotations and role detection is based only on entity annotations positions and classes, not content. Document Filtering doesn't apply because only documents that contain triggers will be further processed.

## 4. Results & Discussion

### 4.1. Experimental setup

We experiment with four base settings for the machine learning method:

- **Pretagged** training and test-sets contain named entities retrieved from our ground truth dataset.
- **PredNEsTest** training-set with pretags from the ground truth, but a test-set with automatically predicted NE-annotations. This represents the practical scenario for our project, but is highly dependent on the quality of the NER model.
- **PredNEs** automatically predicted entities in both training and test-set. We test this setup to see if training on noisy entity mentions improves the models robustness during testing when encountered with similar noise.
- **Plain** no pretagging.

Additionally, we test variants adding the *shortening* augmentation (+Shortening) and *document filter* (+DocFilter). For the rule-based system, we report two setups, one using the ground truth entity mentions and one using automatically predicted entity mentions (analogous to *PredNEsTest*).

### 4.2. Rule-based vs. Machine Learning

Table 2 shows that the machine learning systems significantly outperform our rule-based systems no matter if tags are generated from ground truth information or are automatically predicted. Interestingly, the trained model without any pretagging (*Plain*) still performs similar in role detection compared to a rule-based system working with pretagging information.

In Table 3 the results between the respective best models are shown per category. We observe that the machine learning system is slanted heavily to achieve high precision values. The document filter rule has part in this, reducing recall by appr. one percentage point, but also without filter, a significant slant towards precision remains. Depending on the use of the annotations, this may be problematic. Especially when the annotations are used as a tool to find interesting data points, which are then manually investigated, false positives would likely be less problematic than false negatives.

In a more thorough review of the errors found in the machine learning predictions (specifically *Pretagged+Shortening+DocFilter*) and the rule-based predictions, we observe three main points that the machine learning system is able to handle better:

First, in our dataset, people or organizations represented by someone else are not annotated as taking part in the event (their connection to the event is handled in the form of a relationship

**Table 2**

Micro f1-score with standard deviation for each task: trigger detection, role detection with given trigger and role detection with automatically predicted trigger.

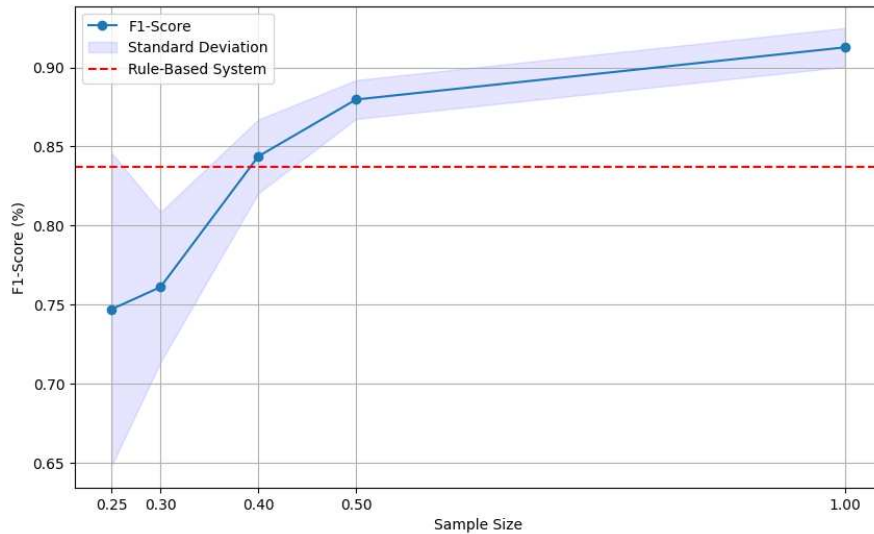
	trigger	roles /w gt trigger	roles /w pred trigger
Rule-based	0.8591 ± 0.0457	0.8674 ± 0.0356	0.8374 ± 0.0442
Pretagged	0.8855 ± 0.0140	0.9112 ± 0.0178	0.8720 ± 0.0115
Pretagged+DocFilter	0.8855 ± 0.0140	0.9151 ± 0.0205	0.8951 ± 0.0072
Pretagged+Shortening	<b>0.9028 ± 0.0261</b>	0.9208 ± 0.0141	0.9048 ± 0.0096
Pretagged+Shortening+DocFilter	<b>0.9028 ± 0.0261</b>	<b>0.9220 ± 0.0140</b>	<b>0.9127 ± 0.0123</b>
Rule-based with PredNEs	0.8586 ± 0.0478	n/a	0.7044 ± 0.0446
PredNEsTest+Shortening+DocFilter	0.8799 ± 0.0256	n/a	0.8118 ± 0.0454
PredNEsTest+DocFilter	0.8717 ± 0.0219	n/a	0.7708 ± 0.0443
PredNEs+DocFilter	0.8414 ± 0.0650	n/a	0.7518 ± 0.0474
Plain	0.8224 ± 0.0709	n/a	0.7107 ± 0.0646

**Table 3**

Performance metrics per label type between rule-based system and machine learning system (with shortening and pretagging). Both using pretagging derived from ground truth data.

	Rule-Based			Machine Learning		
	Recall	Precision	F-Score	Recall	Precision	F-Score
Trigger	80.42%	<b>92.37%</b>	85.91%	<b>89.61%</b>	91.13%	<b>90.28%</b>
Seller	<b>93.96%</b>	76.94%	84.37%	89.69%	<b>95.23%</b>	<b>92.22%</b>
Buyer	81.55%	89.62%	85.37%	<b>88.10%</b>	<b>93.30%</b>	<b>90.54%</b>
Property	83.89%	82.46%	82.96%	<b>84.83%</b>	<b>96.91%</b>	<b>90.45%</b>
Price	79.51%	83.14%	81.18%	<b>87.40%</b>	<b>97.41%</b>	<b>92.00%</b>

between them and the person representing them). E.g. "Es verkauft Hans Vöglin innamen seines bruders kinder" (Hans Vöglin sells in the name of his brothers children...) only classifies "Hans Vöglin" as seller, but not "kinder". Our rule-based system does not contain a rule to ignore these mentions when looking for the roles "seller" and "buyer". Writing rules for these cases isn't trivial either, as phrasing and spelling of words indicating these occurrences varies. The machine learning system was able to correctly ignore these mentions in the examples we investigated manually. Second, as already expected in the methodology section, the rule-based system struggles with the misidentification of buyer as seller, and conversely, seller as buyer. We observe that the machine learning system reduces the amount of errors of this kind by two thirds. Finally, we observe a remarkable difference when it comes to slightly altered phrasing in the documents. While the machine learning system still fails when confronted with completely foreign structures (such as a property purchase being discussed as a past event in the middle of a rent purchase), it can handle small alterations quite well.



**Figure 2:** F1-Score plotted in relation to size of training data for the machine learning system. 134-135 samples in total training+validation (depending on fold).

### 4.3. Learning Curve Analysis

Figure 2 illustrates the performance of the machine learning system (+*Shortening+DocFilter*) compared to the rule-based system. We observe that using around 40% of the training material (appr. 54 samples) will result in role annotations comparable to the rule-based system, while using 50% will achieve significantly better results. As usual for machine learning systems, the increase in performance lessens with increasing sample size.

### 4.4. Impact of Variants

The **Shortening** augmentation improves the scores in all settings where it was applied. The strongest difference could be observed when evaluating roles with predicted triggers (p-value = 0.0131). During error analysis, we found that the removed tokens can also result in a loss of relevant information. Specifically, clauses where a husband is named in conjunction with his wife, e.g. "Es verkaufen Hans, seine Frau Anna..." (Hans and his wife Anna sell...) which would get shortened to "Es verkaufen Hans, Anna...", followed by the names of the sellers, would sometimes result in the misidentification of the wife as a buyer, while the un-shortened model would classify these instances correctly. We thus see the shortening strategy as a success with further research required for more fine-grained variants, e.g. only shortening mentions when certain conditions are met.

The **document filter** rule worked well and improved the results over the board. When shortening is not applied at the same time, we observe a significant improvement (p-value = 0.0316). Otherwise we still observe a positive trend (p-value = 0.0518).



#### 4.5. Practical usability

Our final evaluation of this system in a practical use case scenario is mixed. On one hand, the system produces annotations that may well be used to create data of larger quantities where general trends can be observed. An example for possible analysis could be to combine the role annotations with the nested entity annotation to observe economic interactions between occupational groups over time. On the other hand, the systems show a - larger or smaller, depending on the method applied - amount of bias of only finding the events when the structure of the document fits one of the three main templates. So any conclusions drawn from the predicted event information need to consider this bias with caution.

### 5. Conclusion

In this case study, we've shown that even with relatively simple means, we can achieve automated annotations which are usable in historical research. The scope of this case was intentionally kept small to simplify evaluation and interpretation, but future research in our project will explore how these systems perform across a broader range of event types. Most event types that occur with sufficient frequency for machine learning are of similar structural homogeneity to the documents in this study. Therefore we assume the findings for *property purchases* will also be applicable to other event types. We also aim to explore how transfer learning can benefit event recognition for less frequent event types. We've shown that with our kind of data, a machine-learning system can outperform a rule-based system by a significant margin even when only little training data is available. Writing rules may be quicker than annotating documents still, but considering both systems rely on pretagged texts, the amount of necessary work can probably be reduced significantly if events and entities are annotated at the same time. When working with any data annotated by these methods, knowledge of the bias that is inherent to them is crucial. For example, the samples which do not fit the main templates might be coming from a very specific source, which would lead the automated system to miss most documents from that specific source, which would distort whatever conclusions we're trying to draw from the quantitative results. But this study is only the first foray into event extraction in historical texts and only looked at two quick-to-implement and easily accessible methods. In future research the possible application of LLMs to this task should be investigated, as LLMs have shown to perform well in low-resource scenarios [7], but their applicability to historical German must be evaluated first.

### References

- [1] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota, 2019, pp. 54–59. DOI: 10.18653/v1/N19-4010.

- [2] A. Akbik, D. Blythe, and R. Vollgraf. “Contextual String Embeddings for Sequence Labeling”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA, 2018, pp. 1638–1649.
- [3] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet. “Named Entity Recognition and Classification in Historical Documents: A Survey”. In: *ACM Comput. Surv.* 56.2 (2023). DOI: 10.1145/3604931.
- [4] T. Hodel, I. Prada Ziegler, and C. Schneider. *Pre-Modern Data: Applying Language Modeling and Named Entity Recognition on Criminal Records in the City of Bern*. Presented at the Digital Humanities 2023. Collaboration as Opportunity (DH2023), Graz, Austria. 2023. DOI: 10.5281/zenodo.8107616.
- [5] Q. Li, J. Li, J. Sheng, S. Cui, J. Wu, Y. Hei, H. Peng, S. Guo, L. Wang, A. Beheshti, and P. S. Yu. “A Survey on Deep Learning Event Extraction: Approaches and Applications”. In: *IEEE Transactions on Neural Networks and Learning Systems* 35.5 (2024), pp. 6301–6321. DOI: 10.1109/tnnls.2022.3213168.
- [6] I. Prada Ziegler. *What’s in an entity? Exploring Nested Named Entity Recognition in the Historical Land Register of Basel (1400-1700)*. Presented at the Digital Humanities Benelux 2024, Leuven, Belgium. 2024. DOI: 10.5281/zenodo.11500543.
- [7] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, and G. Wang. *GPT-NER: Named Entity Recognition via Large Language Models*. preprint arXiv 2304.10428. 2023. arXiv: 2304.10428 [cs.CL].

## A. Event Annotation Example

<event:sale> Gend ze <trigger> kaufen </trigger> <seller> Heinrich Trech von Lauezhut der Kremer </seller> u <seller> Margareth Lang Walcherin sin ewirtin </seller> , <buyer> Blesin Winsperg dem schnider </buyer> u . <buyer> Margarethen siner ewirtin </buyer> , <property> daz Hus u . Hofstatt genant zer Thannen , so gelegen als man von dem Vischmergt heruf zem Sunfegen gat [...] , ist erb von dem gotshus Lienh denen jährl darab gand 3 lb 21 lot pfeffer ze wysung </property> um <price> 150 fl . </price> </event:sale>

*appr. english translation: Give to buy Heinrich Trech of Lauezhut the trader and Margareth Lang Walcherin his wife, Blesin Winsperg the taylor and Margareth his wife, the property called zer Thannen, lies when you go from the cattle market to zum Sunfegen [...], is owned by the church St. Lienhart which is paid 3 lb 21 lot of pepper for 150 lb .*

## B. Ground Truth Example With Pretagged Text (BIO-Format)

<b>Token</b>	<b>Role</b>
Gibt	O
ze	O
kaufen	B-Trigger
[B-PER]	B-Seller
Heinrich	O
Trech	O
[E-PER]	O
[B-LOC]	B-Property
daz	O
Hus	O
zer	O
Tannen	O
[E-LOC]	O
um	O
[B-MONEY]	B-Price
150	O
fl.	O
[E-MONEY]	O