Building a Corpus from Handwritten Picture Postcards: Transcription, Annotation and Part-of-Speech Tagging

Kyoko Sugisaki, Nicolas Wiedmer, Heiko Hausendorf
German department, University of Zurich
Schönberggasse 9, 8001 Zurich, Switzerland
{sugisaki, nicolas.wiedmer, heiko.hausendorf}@ds.uzh.ch

Abstract

In this paper, we present a corpus of over 11,000 holiday picture postcards written in German and Swiss German. The postcards have been collected for the purpose of text-linguistic investigations on the genre and its standardisation and variation over time. We discuss the processes and challenges of digitalisation, manual transcription, and manual annotation. In addition, we developed our own automatic text segmentation system and a part-of-speech tagger, since our texts often contain orthographic deviations, domain-specific structures such as fragments, subject-less sentences, interjections, discourse particles, and domain-specific formulaic communicative routines in salutation and greeting. In particular, we demonstrate that the CRF-based POS tagger could be boosted to a domain-specific text by adding a small amount of in-domain data. We showed that entropy-based training data sampling was competitive with random sampling in performing this task. The evaluation showed that our POS tagger achieved a F1 score of 0.93 (precision 0.94, recall 0.93), which outperformed a state-of-the-art POS tagger.

Keywords: postcard corpus, POS tagging, German

1. Introduction

In this paper, we report the construction of the language resource *Ansichtskartenkorpus* (*[anko]*), 'picture postcard corpus', containing over 11,000 holiday postcards written in Standard German and Swiss German. They were manually transcribed and annotated with structural and discourse-related information, and then automatically annotated with text segmentation, lemma and part-of-speech (POS) information.

We will first characterise the texts contained in the resource (Section 2), and then describe their manual transcription and annotation before outlining the development of a NLP toolkit for text segmentation and POS annotation (Section 3).

2. Data Source

The holiday postcards were collected at the University of Zurich from 2009 to present day for the purpose of text-linguistic investigations on the genre and its standardisation and variation over time. The postcards included in our corpus were sent by post from people on holiday, mainly from Switzerland but also from Italy, Germany and other European countries to their family, friends, colleagues and neighbours living in the German-speaking area of Switzerland. About 95% of the cards (11,760 cards) were written mainly in Standard German. The remaining part of the corpus is comprised of postcards written mainly in Swiss German. Although the postcards were dated from 1898 to 2016, the majority were written in the 1980s (22%) and 1990s (19%). On average, a post card contains 50 words, while individual post cards vary from one to 350 words.

3. Corpus Construction

In this section, we describe the process of digitalisation, transcription, and annotation carried out manually and automatically to build the corpus of the collected postcards.

3.1. Overall Pipeline: From Digitalisation to XML with Linguistic Annotation

Because the collected holiday postcards were in paper format, we first scanned the front and back of each card. We then considered using an optical character recognition (OCR) system to extract the texts from the scanned images. However, the postcards were handwritten in German, and OCR systems do not work well for handwritten texts in languages other than English. Therefore, we decided on manual transcription for which we developed a web-based tool. The user interface is illustrated in Figure 3.2. Each scanned card was integrated into the tool. The tool displayed the front and back images of each card on the left side and the transcription and annotation forms on the right side. Thus, the transcribers could directly transcribe handwriting, mark paragraphs, note textual discourse structures (e.g. greetings) and enter metainformation (e.g. dates). The data were then saved in a MySQL database, which we then converted to an XML representation. We then incorporated our automatic annotations in the XML: 1) text segmentation, 2) lemma and 3) POS tags.

3.2. Transcription and Manual Annotation

The picture postcards written in Standard German were transcribed and annotated by four transcribers in a typing office in Germany. The Swiss German postcards were transcribed and annotated by a student whose native language is Swiss German. To ensure the quality of the transcription and the manual annotation, during the process of transcription and annotation, three students checked samples, corrected them manually and gave feedback to the typing office.

Our corpus consisted of the main texts as primary data and textual properties as metadata. A picture postcard consists of two sides – the front side and the back side. The front side of a modern postcard typically includes images of tourist attractions and landscapes, including the name of the location, whereas the back side consists of an address field on the

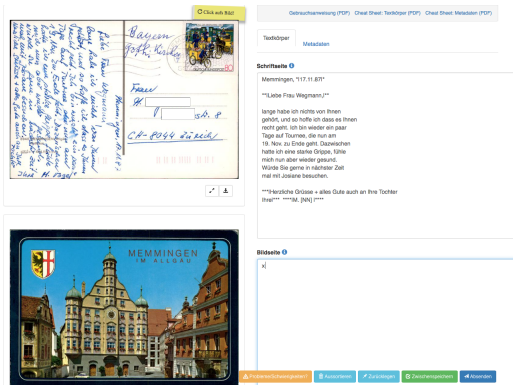

Figure 1: Web-based manual transcription/annotation tool

(A) Word/lemma features	
A1	**Word form**: real word forms
A2	**Normalized word form**: all lower case and without ü
A3	**Character type of unit**: word form is categorised into the following classes: (1) all special characters (2) all numbers (3) capitalized (4) all alphabets without capitalization (5) mixed of all possible character without capitalization
A4-7	**Suffix**: the last 4, 3, 2, 1 character of words, respectively.
A8	**Lemma**: generated by TreeTagger
(B) POS	
B1	**POS**: generated by TreeTagger
B2	**POS**: generated by Stanford POS tagger
(C) Semantic cluster features	
C1-2	**Brown clustering**: Brown clustering is used in 4 digits (D1) and all digits (D2)
C3	**Word2Vec**
C4	**Fasttext**

Table 1: Features for CRF-based POS tagging

right and a message field on the left. During the transcription process, the message field was transcribed and regarded as primary data. The address field (e.g. name, postal code, location and country of the receiver) was considered metadata, including latent information, such as the genders of both receiver and the sender, as well as the presence of sketches drawn by the latter.

In addition, our transcribers annotated textual discourse-related information during the transcription process. The message field of a holiday postcard is generally structured as follows: 1) a preface (date, sometimes location, temperature or weather); 2) a salutation (e.g. *Dear Heidi*); 3) the main message; 4) greeting including closing (e.g. *Cheers*); 5) the signature of the sender. During the transcription, the preface, salutation, greeting and signature were marked directly on the text. Each beginning and end of these discourse zones were marked with unique markdowns. The markdowns consisted of character sequences that hardly appeared in the main text. The salutation was marked as star star bar `**|`, and the closing was marked as `|**`. For example, the salutation *Dear Heidi* in the main text was annotated as `**|Dear Heidi|**`. Hence, minimal annotation was required, and the mapping to XML opening and closing tag was straightforward.

We considered that sensitive data in the corpus should be explicitly coded. The picture postcards often contained private information, such as the name and address of the receiver, the telephone number or even the bank account number of the sender. Therefore, the transcribers did not include such sensitive information but coded as `[Vertraulich]` (i.e. 'confidential') in the message field. In particular, family names are coded as `[NN]` (i.e. the short form of *Nachname* or 'family name'). The sensitive data in the address field were marked as such to ensure that they will not be released in the corpus.[1]

3.3. Automatic Text Segmentation

The primary texts were segmented into paragraphs, sentences and words. The segmented texts were then structured in a XML representation.

Generally in German, punctuation segments a text into sentences, and spaces are used to segment a sentence into words. However, this rule of thumb was not always applicable to the sentence segmentation of the postcards, particularly with regard to the following cases: 1) punctuation was a part of a token with preceding characters; and 2) punctuation was absent. Case 1 refers to abbreviations (e.g. *z.B.* instead of *zum Beispiel* or 'for example') and brand or proper names (e.g. *Sat.1*), which is also common in Standard German orthography. Case 2 refers to freestanding lines, which typically ended with a wide blank space or extra line spacing, and which often omitted punctuation, such as titles, subtitles, addresses, dates, greetings, salutations and signatures (Official German Orthography, 2006). Dates, greetings, salutations and signatures belong to the core text zones of postcards. In addition, freestanding lines were often extended to the end of the paragraph in the texts of the postcards. Furthermore, the following use of punctuations is also common in postcards, which differs from Standard German orthography: (a) repeated punctuation (e.g.,*!!!*,*???*,.....) in order to emphasise words, phrases and sentences; (b) the use of emotional pictograms that are typically composed of punctuation (e.g. *:),;-)*). Based on these peculiarities, we developed a statistical sequential sentence segmentation system that differentiates punctuations into Case (1) and the sentence boundary, and deliberately handles Case (2) (Sugisaki, 2017).

With regard to tokenisation, the texts of the postcards showed a frequent use of contractions, which is also common in internet-based and computer-mediated communication (Bartz et al., 2013). In the contractions, the verb was often combined with the pronoun *es*, 'it', and delimitated by an apostrophe (e.g. *gibt's* instead of *gibt es*, 'gives it'). The apostrophe was sometimes omitted (e.g. *gibts*). Nonetheless, not only verbs are concatenated with the pronoun, but also in 'wh question' words (*wenn's/wo's* instead of *wenn es/wo es*, 'when/where it') and prepositions (*auf's* instead of *aufs* or *auf das*, 'on the'). Based on this observation, we developed a simple rule-based tokeniser in which *'s* was separated from the remaining part of the token if it was not a noun. If it was a noun, the *'s* was considered a genitive marker and part of the token. We used TreeTagger (Schmid, 1995) to obtain the POS information. However, in the case of contractions without apostrophes, TreeTagger does not provide an accurate POS tag. Contractions without apostrophes do not belong to standard orthographies, which might cause this difficulty. We observed that frequently used verbs, such as *give, be* and *have* often occurred with the reduced pronoun *s* without an apostrophe. Therefore, we created a list of these verbs and some wh question words in order to separate *s* from them.

[1] A sample of our corpus will be available at `http://ansichtskartenprojekt.de`

3.4. Part-of-speech Tagging

The segmented tokens were further annotated with POS tags that were integrated into the XML representation. We developed a POS tagger for the postcards. The texts comprised a mixture of Standard German and Swiss German. In addition, the targeted texts were in written form, but conceptually, they were in near-oral language (Koch and Oesterreicher, 2008; Dürscheid, 2016). An off-the-shelf POS tagger is typically trained on a corpus of newspapers written in Standard German. A newspaper article belongs to the category of a prototypical written language in both form and concept. Furthermore, it contains fewer orthographical deviations. Therefore, we experimented with features and training data to determine the best method for optimising the accuracy of the tagger applied to the postcard text in this study.

3.4.1. Experimental Setting

In the experiments, we used the tagging method of conditional random fields (CRF) (Lafferty et al., 2001). CRF is a supervised machine learning method for sequences. For the experiments, we created the following three data sets:

1. TüBa-D/Z v. 10, Tübinger Baumbank des Deutschen/Zeitungskorpus (Telljohann et al., 2012), which is a German newspaper corpus (1.787.801 tokens, henceforth *TüBa*). In our first experiments, approximately 80% of the TüBa (803.040 tokens (henceforth, *TüBa80*) were used as training data, and 20% of the TüBa tokens were used as test data (252.784 tokens, henceforth TüBa20). In the second experiment, we used all the TüBa (*TüBa100*) tokens as training data. In addition, we used a cross-validation data set (2.239 tokens) in all experiments.

2. NOAH's Corpus of Swiss German Dialects (henceforth, *NOAH*) (Hollenstein and Aepli, 2014) is a Swiss German corpus (94.306 tokens) that contains a variety of texts (blogs, reports, Wikipedia, etc.). In our experiments, we used the corpus as training data.

3. From the Ansichtskartenkorpus, or 'picture postcard corpus'(henceforth, *ANKO*), we first manually annotated 200 postcards to derive the test data. The test data were sampled randomly from the corpus and divided into two sets: 100 cards for the experiment (5.048 tokens, henceforth *ANKO-TEST*) and 100 cards for the evaluation (5.341 tokens, henceforth *ANKO-EVAL*). In addition, we manually annotated 1,500 sentences for the experiments. The sentences were used as training data, and they were sampled in three ways: 1) 300 sentences were selected randomly (henceforth, *ANKO-R*); 2) 1,200 sentences were selected based on word 4-gram-based entropy scores according to four measurements. We describe the entropy sampling method in Section 3.4.3. In our experiments, we used the Standard German sub-corpus of *ANKO*.

The set of linguistic features used in our experiments is provided in Table 1. The features were divided into (A) word and lemma, (B) POS features generated by the POS tagger TreeTagger (Schmid, 1995) and the Stanford POS Tagger (Toutanova et al., 2003); and (C) semantic clusters generated by unsupervised machine learning methods, that is, Brown clustering[2] (Brown et al., 1992), [3] (Mikolov et al., 2013) and fasttext[4] (Bojanowski et al., 2016).
In the following subsections, we describe the experiments using the set of linguistic features and the data sets.

3.4.2. Features

We trained CRF models on the training set of TüBa and tested them on the test set of TüBa and ANKO. We trained four different types of features (A to C in Table 1) separately and all features in context window 0 (i.e., current tokens). The results are shown in Table 2. As expected, tagging accuracy (F1 score) was lower if the training data and test data were derived from different domains. Regardless of the test data, the best features were the word and lemma features (A). The morphosyntactic analysis using the existing POS taggers showed a lower performance, and the semantic features (B) did not achieve high accuracy. However, the combination of these three types of features outperformed the word/lemma features. We extended the feature sets of (A), (B) and (C) from context window 0 (current tokens) to 5 left and right context windows. The results are shown in Table 1. The main finding was that the window side did not affect the accuracy as much as expected. However, the wider context window size slightly improved the accuracy of the test set of TüBa. Therefore, we conducted further experiments using the combination of the feature sets (A), (B) and (C) in context windows 0 to 5.

3.4.3. Training Data

In this section, we investigate the following challenges: 1) how to boost the tagging accuracy in texts with mixed languages and 2) whose domain and morphosyntactic distribution were different from newspapers.
To handle the first challenge, we added the Swiss German training data, NOAH. The results are shown in Table 3. The addition of the Swiss German training data produced results that were similar to those of the model that was trained only on TüBa100, but it did not improve the tagger.
To address the second challenge, we added small amounts of five types of training data from ANKO to the TüBa100 and NOAH training data. The first in-domain training data were randomly selected from ANKO. The remaining data sets were selected using a cross entropy score. Cross entropy is a variant of perplexity that is used to compare different probability models. The score is measured as follows (Jurafsky

[2]For Brown clustering, we used the implementation of P. Liang. To create 100 clusters, we trained the model on TüBa100, NOAH, ANKO (normalized word form). The first 4 digits and all digits are used as features.

[3]For word2vec, we used gensim with parameters skip-gram, 500 dimensions, context window 5. For K-means clustering, we used the scikit-learn to create 30 clusters.

[4]We used the fasttext with parameters, CBOW, 200 dimensions, context window 5, 5 word ngrams. For K-means clustering, we used the scikit-learn to build 20 clusters.

Context window	0				0-1	0-3	0-5
Feature	**Feature A**	**Feature B**	**Feature C**	**Feature A-C**			
TüBa-Test	.968 (.968,.968)	.960 (.960,.961)	.893 (.893,.894)	.974 (.974,.974)	.977 (.977,.977)	**.978 (.978,.978)**	**.978 (.978,.978)**
ANKO-Test	.883 (.886,.881)	.848 (.850,.846)	.795 (.796,.794)	**.897 (.900,.895)**	.895 (.897,.893)	.892 (.895,.890)	.895 (.897,.893)

Table 2: Experiments with features in context window 0, 0-1, 0-3, 0-5: Training data =TüBa80:F1 score (precision, recall)

(2009, pp. 117)):

$$H(w_1 \ldots w_n) = -\frac{1}{N} logP(w_1 \ldots w_n) \quad (1)$$

The goal of the in-domain training data selection was the automatic selection of a small number of in-domain sentences that might improve the tagging accuracy. Ideally, the in-domain sentences to be selected were not observed in the training in TüBa and NOAH but were typical in ANKO. We considered two methods: 1) ranking-based entropy scoring (henceforth, method [A]) and 2) difference-based entropy scoring (henceforth, method [B]). Ranking-based entropy scoring is a measurement of how informative in-domain sentences are based on a language model trained on out-of-domain data. The entropy scores were ranked in order from high to low. In this method, in-domain sentences with high entropy scores were assumed distinct from the out-of-domain data and thus more informative. This method is compatible with Axelrod and Gao (2011) in which perplexity was used instead of cross entropy. We inspected the top 300 sentences. They included salutations, greetings, signatures and dates. These discourse types are typical in postcards but are rarely included in a newspaper corpus. In contrast, difference-based entropy is a measurement of differences in entropy scores based on a language model trained on both out-of-domain and in-domain sentences. In-domain sentences were considered informative if the difference in score was large. This method is based on Moore and Lewis (2010). We inspected the top 300 sentences. These sentences were similar to those selected by the ranking-based entropy scores, and they were a mixture of typical discourse structures.

However, the selected sentences did not include in-domain interpersonal and fragmental sentence patterns typically used in private communication. Thus, we did not find any sentences whose subject was in the first or second person, such as *Danke für Deine Karte.* ('Thank you for your card') or fragments such as *sind glücklich hier oben gelandet* ('happily landed up here above'). Here, we found that the variance in higher entropy scores was high in TüBa (mean: 10, variance: 914) and low in ANKO (mean: 1, variance: 3), which indicated that the difference-based entropy scores were mainly guided by the TüBa scores. Therefore, these two methods selected similar sentences.

To detect typical main sentences in ANKO, we introduced two methods: in-domain ranking-based entropy score (henceforth, method [C]) and a difference-ranking-based entropy score (henceforth, method [D]). Method (C) was used to select the sentences with lowest entropy scores based on a language model trained on the in-domain data. In method (D), we simply ranked the entropy scores trained on TüBa and on ANKO, and we ordered the difference in ranking from high to low.

Training data	Test
TüBa100	.899 (.902,.897)
TüBa100 + NOAH	.898 (.901,.896)
TüBa100 + NOAH + 100 ANKO-A	.910 (.913,.908)
TüBa100 + NOAH + 100 ANKO-B	.908 (.911,.906)
TüBa100 + NOAH + 100 ANKO-C	.922 (.924,.920)
TüBa100 + NOAH + 100 ANKO-D	.926 (.928,.924)
TüBa100 + NOAH + 100 ANKO-R	.931 (.934,.929)
TüBa100 + NOAH + 300 ANKO-R/A/B/C/D	.941 (.943,.939)

Table 3: Experiments with training data with features (A), (B), and (C), and test on the *ANKO-TEST*: F1 score (precision, recall)

We experiment on these domain-data selection methods (A)-(D) with random selection (R) as our baseline. For that, we manually annotated 300 sentences for the training set (R) and (A)-(D). The results are shown in Table 3. The 300 sentences selected by the (D) method outperformed the other three entropy-based sampling methods, which indicated that ranked-difference-based entropy scoring is a viable sampling method, particularly if differences in the variance of the entropy scores between out-of-domain and in-domain data are large. However, the selected sentences did not outperform the in-domain data that were selected at random. Finally, we tested the models trained on TüBa, NOAH and 1,200 training sentences in the postcards, which achieved the best F1 score of 0.94.

3.4.4. Evaluation

To evaluate the developed POS tagger, we created a test set that was derived from the postcard corpus (ANKO-EVAL). We re-trained the CRF model with the features A, B, and C and the training data, TüBa100, NOAH, ANKO (i.e. ANKO-Test, all ANKO in-domain training sentences R/A/B/C/D). For the comparison, we used TreeTagger. The evaluation revealed that our POS tagger achieved a F1 score of 0.93 (precision 0.94, recall 0.93), which outperformed TreeTagger's F1 score of 0.86 (precision 0.86, recall 0.86).

4. Conclusion

In this paper, we described the process of digitalising, transcribing and annotating of over 11,000 handwritten postcards. In particular, we demonstrated that the POS tagger could be boosted to a domain-specific text by adding a small amount of in-domain data. We showed that entropy-based training data sampling was competitive with random sampling in performing this task. In future work, we will test our POS tagger on text that is written in Swiss German.

5. Acknowledgments

This work has been funded under SNSF grant 160238. We thank all the project members, Joachim Scharloth, Noah Bubenhofer, Selena Calleri, Maaike Kellenberger, David Koch, Marcel Naef, Dewi Josephine Obert, Jan Langenhorst, Michaela Schnick.

6. Bibliographical References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 355–362.

Bartz, T., Beißwenger, M., and Storrer, A. (2013). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *JLCL*, 28:157–198.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Dürscheid, C. (2016). *Einführung in die Schriftlinguistik*. Vandenhoeck & Ruprecht, Göttingen, 5 edition.

Hollenstein, N. and Aepli, N. (2014). Compilation of a Swiss German dialect corpus and its application to PoS tagging. *COLING 2014*, page 85.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Education International, Upper Saddle River, New Jersey, 2nd edition.

Koch, P. and Oesterreicher, W. (2008). Mündlichkeit und Schriftlichkeit von Texten. In Nina Janich, editor, *Textlinguistik: 15 Einführungen*. G. Narr.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML)*, pages 282–289.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers (ACLShort)*, pages 220–224.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop.*, Dublin, Ireland.

Sugisaki, K. (2017). Word and sentence segmentation in german: Overcoming idiosyncrasies in the use of punctuation in private communication. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*.

Telljohann, H., Hinrichs, E. W., Sandra, K., Heike, Z., and Kathrin, B. (2012). Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Technical report, Universität Tübingen.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03)*, pages 173–180.