Published in: Schmidt, Sara/Siemens, Ray/Kumar, Amit/Unsworth, John (eds.): Digital Humanities 2007.
Conference Abstracts - Urbana-Champaign: University of Illinois, 2007. pp. 166-170.

Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert

1 Introduction

Though XML-annotated text collections are commonplace in humanities computing, the value of the annotation is often underestimated, as interesting applications can be realised by ignoring the content and considering the annotation exclusively. At the same time, the distribution of text collections (e. g., linguistic resources) is often restricted by rigid licence agreements. Usually, a corpus consists of a source text collection (STC) acquired from third parties such as web sites or publishers, and annotation layers that refer to, for example, structural or linguistic properties. In practically all cases the STC is a copyrighted property, so that it is up to the copyright holder to decide if, and under which conditions, the corpus - a crucial part of which is the STC - can be made available to the public or to the research community.

The example we use in this paper is TüBa-D/Z ("Tübingen Treebank of Written German" (Telljohann et al, 2004 & Telljohann et al, 2006)). This manually annotated treebank is based on a CD ROM that contains an archive of the issues the newspaper *die tageszeitung (taz)* has published since 1986. If a researcher (the licencee) wants to obtain TüBa-D/Z, available for academic purposes free of charge, he or she has to sign a licence agreement with Tübingen University's Linguistics Department (the licencer) which states that the licencer is the copyright holder of the annotation and that the STC, as published on the *taz* CD ROM, is copyrighted by contrapress media GmbH. The licencee has to certify that he, she or the institution the person works for has a valid licence for this CD ROM.[1]

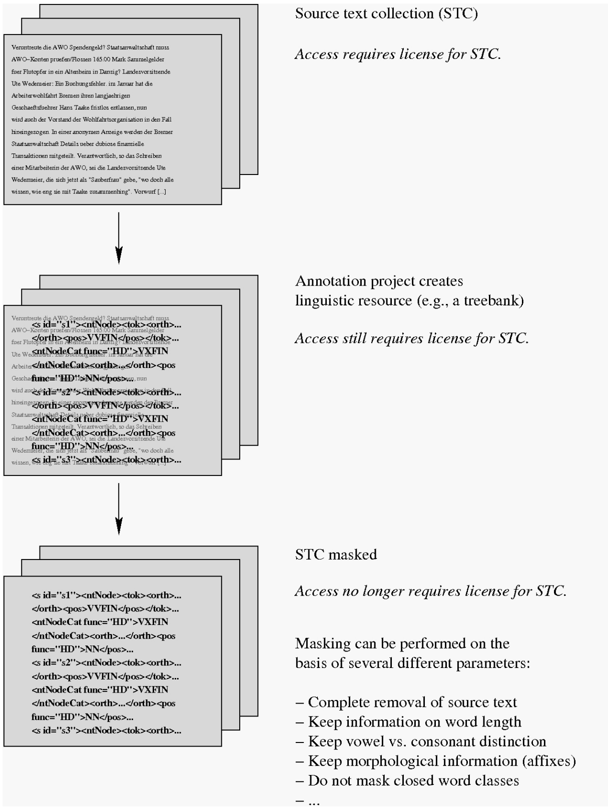

Veruntreute die AWO Spendengeld? Staatsanwaltschaft muss
AWO–Konten pruefen/Flossen 165.00 Mark Sammelgelder
fuer Flutopfer in ein Altenheim in Danzig? Landesvorsitzende
Ute Wedemeier: Ein Buchungsfehler. im Januar hat die
Arbeiterwohlfahrt Bremen ihren langjaehrigen
Geschaeftsfuehrer Hans Taake fristlos entlassen, nun
wird auch der Vorstand der Wohlfahrtsorganisation in den Fall
hineingezogen. In einer anonymen Anzeige werden der Bremer
Staatsanwaltschaft Details ueber dubiose finanzielle
Transaktionen mitgeteilt. Verantwortlich, so das Schreiben
einer Mitarbeiterin der AWO, sei die Landesvorsitzende Ute
Wedemeier, die sich jetzt als "Sauberfrau" gebe, "wo doch alle
wissen, wie eng sie mit Taake zusammenhing". Vorwurf [...]
Source text collection (STC)
Access requires license for STC.
<s id="s1"><ntNode><tok><orth>...
</orth><pos>VVFIN</pos></tok>...
<ntNodeCat func="HD">VXFIN
</ntNodeCat><orth>...</orth><pos
func="HD">NN</pos>...
<s id="s2"><ntNode><tok><orth>...
</orth><pos>VVFIN</pos></tok>...
<ntNodeCat func="HD">VXFIN
</ntNodeCat><orth>...</orth><pos
func="HD">NN</pos>...
<s id="s3"><ntNode><tok><orth>...
Annotation project creates
linguistic resource (e.g., a treebank)
Access still requires license for STC.
STC masked
Access no longer requires license for STC.
Masking can be performed on the
basis of several different parameters:
– Complete removal of source text
– Keep information on word length
– Keep vowel vs. consonant distinction
– Keep morphological information (affixes)
– Do not mask closed word classes
– ...

Figure 1: Masking linguistic corpora by example of the TüBa-D/Z treebank

We propose the notion of corpus masking, i. e., obfuscating the STC, but not the annotation layer(s), the STC is "removed", so that the original licensing restrictions no longer hold for the "new" resource. The advantage is that the valuable annotation information can be made available for free (see figure 1).[2]

2 Corpora – Licence Restrictions – Sustainability

When linguists have created a corpus it can become quite difficult to gain access to the corpus once the project is finished. In an ideal world, academics can turn to a sustainability initiative in order to archive their datasets and to make them available to other researchers, e. g., by means of a web-based corpus platform (Dipper et al, 2006 & Schmidt et al 2006). Apart from issues such as providing standardised markup languages and metadata sets (Chiarcos et al, 2006 & Wörner et al, 2006), sustainability initiatives have to take the copyright of the original data into account.

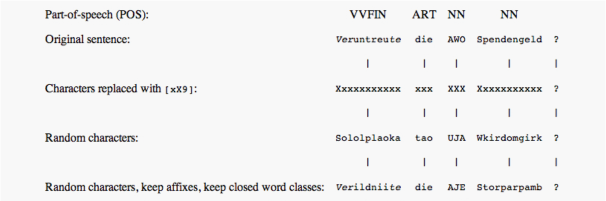
We developed a tool that is able to mask corpora on the fly. Should someone who is interested in a corpus that is available under a rigid licence model not have a valid STC licence, he or she can still receive the corpus, albeit in masked form. A corpus potentially can be associated with *several* accessibility regulations: full access to TüBa-D/Z requires a licence for the *taz* CD ROM, whereas masked versions can be placed under, say, the GNU Free Documentation or a Creative Commons Licence. Therefore, a sustainability initiative has to come up with a flexible system of representing the relationships and dependencies between the STC and the different annotation layers and their individual licence restrictions.

3 How to Mask Linguistic Resources

The easiest option to obfuscate an annotated corpus is to remove the text. A less radical solution substitutes every STC character with, for example, "x" and every digit with "0". In addition to preserving word length, this process retains information on upper and lower case by substituting these with "x" and "X" (Toms & Campbell, 1999).

We developed *CorpusMasker*, a Java-based tool for the parameterised masking of linguistic resources represented as XML documents. The XML element(s) or attribute(s) that comprise the actual words or tokens to be masked (in case of TüBa-D/Z, the `<orth>` element) can be specified to handle arbitrary annotation schemes. CorpusMasker features a dictionary approach: after collecting all word forms, every word is mapped onto a randomly generated string and replaced by that string. Word length can be retained, as well as information on the distribution and positioning of vowels and consonants. If a word is usually written with an initial lower case character and that word appears with an initial upper case character, the same randomised word is used (e. g., "dort" -> "kulp", "Dort" -> "Kulp"). CorpusMasker performs an affix analysis that is similar to morphology induction. The algorithm analyses certain words, masks the roots, but retains the affixes, so that the text is masked but valuable linguistic information that in itself is insufficient to reconstruct the source text or even to interpret the masked text, is kept intact for further analysis. Parameterised masking can be performed with several different degrees of retaining linguistic information, from the complete removal of the STC to a rather light but sufficient masking that keeps, e. g., closed word classes unchanged (see table 1; affixes are marked in italics).[3]

Linguistic corpora often contain POS information so that the randomisation process results in a list that could act as a key to unlock the masked corpus, i. e., to reconstruct the STC. As publication of this complete list would contradict the purpose of the tool, we will only provide a reduced version of the file so that the randomly generated words can be mapped onto POS tags.



Part-of-speech (POS):	VVFIN	ART	NN	NN	
Original sentence:	*Veruntreute*	die	AWO	Spendengeld	?
	|	|	|	|	|
Characters replaced with `[xX9]`:	Xxxxxxxxxxx	xxx	XXX	Xxxxxxxxxxx	?
	|	|	|	|	|
Random characters:	Sololplaoka	tao	UJA	Wkirdomgirk	?
	|	|	|	|	|
Random characters, keep affixes, keep closed word classes:	*Verildniite*	die	AJE	Storparpamb	?

Table 1: Masking examples for "Veruntreute die AWO Spendengeld?"

4 Masked Corpora: What are They Good for?

Our original goal had been to give researchers interested in TüBa-D/Z the option of examining the annotation without ordering the *die tageszeitung* CD ROM first. As our sustainability platform will give access to copyrighted corpora, we will implement the option of masking a corpus archive before every single download to enhance security. Furthermore, a password protected dictionary lookup could be provided that enables researchers to retrieve a small amount of translations from randomised strings back to original words. Following, we sketch some application scenarios for masked corpora.

Unlexicalised parsing A masked corpus can be used for all sorts of unlexicalised training. Charniak (1996) shows that an unlexicalised PCFG trained on treebank annotations is compatible with other unlexicalised parsers. In addition to the masked training data, a minimal amount of testing data was required. In the case of TüBa-D/Z this subcorpus could consist of randomly shuffled example sentences from the treebank with unmasked text and full annotation. Hinrichs et al. (2005) discuss experiments in memory-based learning of anaphora resolution. Their tool is trained on the annotation of TüBa-D/Z and does not take lexical information into account. The features refer to morphological properties, parts-of-speech, syntactic boundaries and grammatical functions, all of which are available in the annotation. In this case even the test data could be generated directly from the masked resource since the annotation includes marking of equivalence classes comprising pronouns and noun phrases. The gold standard for testing consists of these equivalence classes only in which the words are represented by positional indices. The evaluation would then test whether the relevant indices are grouped together correctly. A comparable tool trained on masked corpus data could as well be applied to `real' German texts.

Qualitative and quantitative analyses TüBa-D/Z's annotation can be used for qualitative and quantitative analyses, it includes both syntactic categories as well as grammatical functions. A linguist can, for example, examine which categories occur as predicatives (element PRED). In addition to this qualitative investigation, the corpus also allows a quantitative analysis: what percentage of predicatives is realised by a noun phrase, what percentage is realised by an adjectival phrase or by a prepositional phrase? To give a second example, coordinate structures are marked with the label KONJ; even without knowledge of the word level the treebank annotation gives sufficent information to examine parallelism effects with respect to the structure of the conjuncts: syntactic categories, grammatical functions, modifiers, and length, see, e. g., Levy (2004), and Steiner (2006).

Teaching linguistics and computational linguistics The masked version of TüBa-D/Z contains an unnatural language that acts like German syntaxwise, but the lexicon of this language contains, for the most part, random strings and associated POS tags. This fact makes the masked treebank a valuable resource in the context of teaching computational linguistics. If students have to work with a language that has a known syntax and a rudimentary morphology but lexical entries that bear no meaning whatsoever, they might be able to concentrate better on the tasks of developing grammar rules or improving parsing efficiency (e. g., with regard to unlexicalised parsing). This approach of blanking out semantics is compatible with Chomsky's notion of language as processing a set of symbols.[4]

Evaluating NLP software Another promising application scenario is the evaluation of NLP software. Most tools use n-gram language models, more sophisticated applications can be trained on annotated corpora. With a masked resource it is possible to measure the influence syntactic annotations have concerning precision and recall, as the performance data of an NLP tool with regard to original, as well as slightly and fully masked corpora can be compared. This approach could result in substantial arguments in favour, or against the use of treebanks for training NLP tools.

5 Related Work

Anonymisation methods remove proper nouns and other identity-revealing phrases to protect the privacy of the people mentioned in a text (for example, medical or legal records (Corti et al, 2006, Medlock, 2006, Poesio et al, 2006, & Rock, 2001)). A second application area is concerned with the removal of cues that might reveal the identity of the author of a text. A third area concerns the masking, or obfuscation of texts, as described in the present paper; we are not aware of similar approaches to the masking of linguistic resources.[5]

6 Concluding Remarks

We call our approach parameterised masking because the randomisation process can be influenced with regard to several parameters, so that, for example, certain word classes are not randomised. Typically, when closed word classes such as determiners and prepositions are kept intact, at least part of the original meaning of a sentence can be guessed. This leads us to a crucial question: what happens if we choose to mask only a small number of words (for example, only proper nouns)?

Do we have to mask a certain percentage of words, in order to bypass the STC's licensing restrictions? When does a text that has been masked only minimally become the original text again, so that the licence restrictions *prohibited* the distribution of the pseudo-masked linguistic resource?

Acknowledgements

The authors would like to thank Timm Lehmberg (Hamburg) and Felix Zimmermann (Hannover) for valuable comments with regard to legal aspects of the approach presented in this contribution. Furthermore, we would like to thank our colleague Holger Wunsch for valuable discussions.

Bibliography

Charniak, E. "Tree-bank Grammars." *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96).* MIT Press, 1996. 1031-1036.

Chiarcos, Christian, Timm Lehmberg, Georg Rehm, and Andreas Witt. *Regulating Access to the Sustainability Platform. Technical report.* SFB 441 (Tübingen University), 2006.

Corti, L., A. Day, and G. Backhouse. "Confidentiality and Informed Consent: Issues for Consideration in the Preservation of and Provision of Access to Qualitative Data Archives." *Forum: Qualitative Social Research* 1.3 (2000).

Dipper, S., E. Hinrichs, T. Schmidt, A. Wagner, and A. Witt. "Sustainability of Linguistic Resources." *Proceedings of the LREC 2006 Workshop Merging and Layering Linguistic Information, Genoa, Italy.* Ed. E. Hinrichs, N. Ide, M. Palmer and J. Pustejovsky. 2006. 48-54.

Hinrichs, E., K. Filippova, and H. Wunsch. "What Treebanks Can Do For You: Rule-based and Machine-learning Approaches to Anaphora Resolution in German." *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT 2005), Barcelona, Spain.* Ed. M. Civit, S. Kübler and Ma. Antònia Martí. 2005. 77-88.

Levy, Roger. Presented at the Department of Linguistics, University of Colorado-Boulder, March 11, 2004. 2004.

Medlock, B. "An Introduction to NLP-based Textual Anonymisation." *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.* Ed. N. Calzolari, K. Choukri, A. Gangemi, J. Mariani, B. Maegaard, J. Odjik and D. Tapias. 2006. 1051-1056.

Piez, Wendell. "Way Beyond Powerpoint: XML-driven SVG for Presentations." *Proceedings of XML 2004, Washington, November 2004. IDEA.* Ed. Lauren Wood. 2004.

Poesio, M., M. A. Kabadjov, P. Goux, U. Kruschwitz, E. Bishop, and L. Corti. "An Anaphora Resolution-Based Anonymization Module." *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.* Ed. N. Calzolari, K. Choukri, A. Gangemi, J. Mariani, B. Maegaard, J. Odjik and D. Tapias. 2006. 1191-1193.

Rock, F. "Policy and Practice in the Anonymisation of Linguistic Data." *International Journal of Corpus Linguistics* 6.1 (2001): 1-26.

Schmidt, Thomas, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. "Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources." *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art, East Lansing, Michigan.* 2006.

Steiner, Ilona. "Coordinate Structures: On the Relationship between Parsing Preferences and Corpus Frequencies." *Pre-Proceedings of the International Conference on Linguistic Evidence 2006, Tübingen, February 2006.* 2006.

Telljohann, Heike, Erhard Hinrichs, Sandra Kübler, and Heike Zinsmeister. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report.* Tübingen: Seminar für Sprachwissenschaft, Universität Tübingen, 2006.

Telljohann, Heinke, Erhard Hinrichs, and Sandra Kübler. " The TüBa-D/Z Treebank – Annotating German with a Context-Free Backbone." *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, 2004.* 2004.

Toms, E. G., and D. G. Campbell. " Genre as Interface Metaphor: Exploiting Form and Function in Digital Environments." *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32).* IEEE Computer Society, 1999.

Varga, D., P. Halácsy, A. Kornai, V. Nagy,, L. Németh, and V. Trón. "Parallel Corpora for Medium Density Languages." *International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria, 2005.* Ed. G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov and N. Nikolov. 2005. 590-596.

Wörner, Kai, Andreas Witt, Georg Rehm, and Stefanie Dipper. "Modelling Linguistic Data Structures." *Proceedings of Extreme Markup Languages 2006, Montréal, Québec, August 2006.* Ed. B. T. Usdin. 2006.

[1] The *taz* CD ROM costs about 50 Euros. Licences for other corpora are often more expensive.

2. The institution that created the annotation holds its copyright and can decide the distribution conditions. As modern corpora may comprise several annotation layers created by more than one research group, each group can be considered the creator of its annotation layer and can decide its terms of distribution (as a consequence, every annotation layer should potentially comprise a complete metadata record). Commercially available software tools that were used in the annotation process might restrict the terms of distribution of the resulting data set as well.
3. After DH 2007, a downloadable version of CorpusMasker will be available on our web site under an Open Source licence (`<http://www.sfb441.uni-tuebingen.de/c2/>`).
4. For centuries, typographers and graphic designers use the "Lorem ipsum dolor sit amet" text fragment to evaluate new layouts without resorting to writing actual text. The blind text gives the impression of a natural distribution of characters and whitespace without distracting the reader by conveying any meaning that could be interpreted intuitively. This approach might be useful for visualising masked corpora by means of XML to SVG transformations (Piez, 2004).
5. In a message posted to Corpora-List on Aug 19th, 2006, Péter Halácsy suggested an interesting method to distribute a copyrighted corpus under "fair use" conditions. Part of the copyright notice Halácsy et al. apply to the Creative Commons-based licence of the "Hunglish" corpus (D. Varga et al, 2005) reads: "We prevented the illegal use of copyrighted material by shuffling the texts at sentence level. This form is still useful for research purposes, while it does not infringe upon the rightholders' interests."