Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications

Georg Rehm[1], Andreas Witt[1], Heike Zinsmeister[2], Johannes Dellert[1]

Tübingen University[1]
SFB 441: Linguistic Data Structures

Heidelberg University[2]
Dept. of Computational Linguistics

Abstract

The distribution of linguistic resources such as treebanks and other corpora is often restricted by rigid license agreements. We present a theoretical framework and an implemented tool for the masking of linguistic resources, i. e., an approach to obfuscate or to hide the original primary data (copyrighted text) to enable the free distribution of a linguistic resource as well as additional application scenarios.

1 Introduction

The distribution of linguistic resources is often restricted by rigid license agreements.[1] A treebank or any other type of corpus consists of two parts: (a) one or more source texts, and (b) one or more layers of annotation that refer to linguistic properties of the texts. Usually, the linguistic properties are annotated manually by academics or automatically by software tools; the source text collection (STC) has been acquired beforehand from third parties such as web sites or publishing houses.[2] In practically all cases the STC is a copyrighted property that is subject to access restrictions. At the end of the day it is up to this copyright holder to decide if, and under which conditions, the linguistic resource – a crucial part of which is the STC – can be made available to the public or research community.

The manually annotated treebank TüBa-D/Z ("Tübingen Treebank of Written German", see Telljohann et al., 2004, 2006)[3] is based on a commercially available CD ROM that contains an archive of all the issues of

[1]The authors would like to thank Timm Lehmberg (Hamburg) and Felix Zimmermann (Passau) for valuable comments with regard to legal aspects of our approach. Furthermore, we would like to thank Holger Wunsch (Tübingen) for valuable discussions.

[2]The source text collection might also consist of transcribed spoken language, in which case similiar problems emerge with respect to the privacy of the speakers (see section 5).

[3]Throughout the paper we will come back to this treebank as an example. TüBa-D/Z currently consists of ca. 27,000 sentences (470,000 tokens). It comprises annotation of, among others, parts-of-speech, syntactic constituency, and grammatical functions.

the newspaper *die tageszeitung* (*taz*) that have been published since 1986. If a researcher (the licensee) wants to obtain TüBa-D/Z, available for academic purposes free of charge, he or she has to sign a license agreement with the Linguistics Department at Tübingen University (the licenser). The agreement states that the licenser is the copyright holder of the linguistic annotation and that the STC, as published on the CD ROM, is copyrighted by the company contrapress media GmbH. Therefore, the licensee has to sign a statement that certifies that he or she or the institution the person works for has a valid license of this CD ROM; furthermore, a copy of the CD ROM invoice has to be submitted as additional proof.[4] Only if the licenser receives the signed agreement and a copy of the invoice, the licensee can be sent the access information for the password-protected TüBa-D/Z download site.

This article introduces an approach that we call the masking of linguistic resources, in order legally to bypass licensing restrictions such as the ones described in the previous paragraph. The idea is to mask the STC, but not the layers of linguistic annotation. This approach practically removes the STC, so that the original licensing and copyright restrictions no longer hold for the new resource. The advantage is that the information that is most crucial and most interesting to other linguistics researchers, the annotation itself, can be made available *without* any restrictions (see figure 1 in Rehm et al., 2007b, p. 166).[5] We think that our approach is especially valuable for corpora comprising syntactic annotation including phrase structure information. Such treebanks normally offer token-related annotation such as part-of-speech tags as well as hierarchical annotation structures beyond the word level. In this scenario, masking the word tokens leaves information rich enough to be used independently.

Section 2 discusses the masking of linguistic corpora in sustainability projects. Our software tool, CorpusMasker, is described in section 3. Section 4 highlights application scenarios in which masked corpora can be used in a practical way. Section 5 addresses related work.

[4]The *die tageszeitung* CD ROM costs about 50 Euros. Licenses for other (newspaper) corpora are often, if available at all, much more expensive.

[5]The institution that created the linguistic annotation is the copyright holder of the annotation. Therefore, it is up to this institution to decide the conditions under which the now masked linguistic resource is to be made available to third parties. Usually, the aim is to make the resource available online at no cost. To complicate matters even further, modern corpora may be comprised of *multiple* annotation layers that have been 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 (this circumstance has very serious consequences for the annotation of metadata: not only the complete corpus, but every single annotation layer should potentially comprise a complete metadata record). Commercially available software tools that were used in the annotation process (for example, POS taggers) might restrict the terms of distribution of the resulting data set as well.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

2 Corpora – License Restrictions – Sustainability

It is the goal of linguistic sustainability initiatives to archive and to make available heterogeneous sets of linguistic resources, i. e., not only corpora but also linguistic software, so that interested parties are able to access them (Dipper et al., 2006). Nowadays researchers predominantly work with empirical data, they use and they create corpora, normally with a linguistic theory and a specific research question in mind. When a project is finished it can be very difficult to gain access to the corpus. In an ideal world, academics can turn to a sustainability initiative (sometimes also referred to as preservation projects) in order to archive their datasets (Trilsbeek and Wittenburg, 2006) and to make the data available to other researchers, e. g., by means of a web-based corpus repository. Apart from the obvious issues such as providing comprehensive and standardised markup languages and metadata specifications (Schmidt et al., 2006), sustainability initiatives need to take extra care of respecting the copyright of the original data (for details see Lehmberg et al., 2007, 2008, Zimmermann and Lehmberg, 2007, Newman, 2007). When an academic or a research institution is interested in uploading their treebank, the web-based platform must be able to restrict access to the data if needed. From the point of view of the sustainability initiative as well as the original supplier of a corpus, it would be advantageous to bypass the licensing restrictions for several reasons, such as enlarging the potential audience of a data collection and extending the visibility of the sustainability initiative within the community (see section 4).

There are two aspects of corpus masking within the context of sustainability initiatives that we would like to emphasise. First, we developed a tool that is able to mask corpora on the fly. The tool can be integrated into a web-based corpus delivery platform. Should someone who is interested in a corpus that is available under a license model as described in section 1 not have a valid license for the STC, he or she can still receive the corpus, albeit in masked form. Second, a linguistic corpus potentially can be associated with *several* accessibility regulations. For example, full access to the TüBa-D/Z treebank requires the licensee to have a valid license of the *taz* CD ROM, whereas the masked version of TüBa-D/Z can be placed under, say, the GNU Free Documentation License. As a consequence, a sustainability initiative has to come up with a flexible system of representing the relationships and dependencies between the source texts and the different layers of annotation and their corresponding license restrictions: if one or more layers whose license regulations are very restricted, are removed from a corpus that is about to be delivered, the next restrictive license of the remaining part of the corpus needs to be applied. This representation should be included in the metadata records of any corpus and a corresponding process logic should be integrated into the platform (Rehm et al., 2007a).

3 How to Mask Linguistic Resources

There are several ways to mask a corpus, i. e., to obfuscate the texts a corpus is made up of. The most simple option is completely to remove the textual content of the collection. A slightly less radical solution substituted every single character contained in a word of the STC with one specific character such as “x” and every digit with, for example, “0”. Next to preserving information on the length of a word, this process could preserve information on upper and lower case characters by substituting capital letters with “X” and lower case characters with “x” (Toms and Campbell, 1999). Additional mappings can be defined in a step-by-step manner, so that more and more information related to the STC can be retained (the realisation of this process weakens the aim that is responsible for masking a text, though).

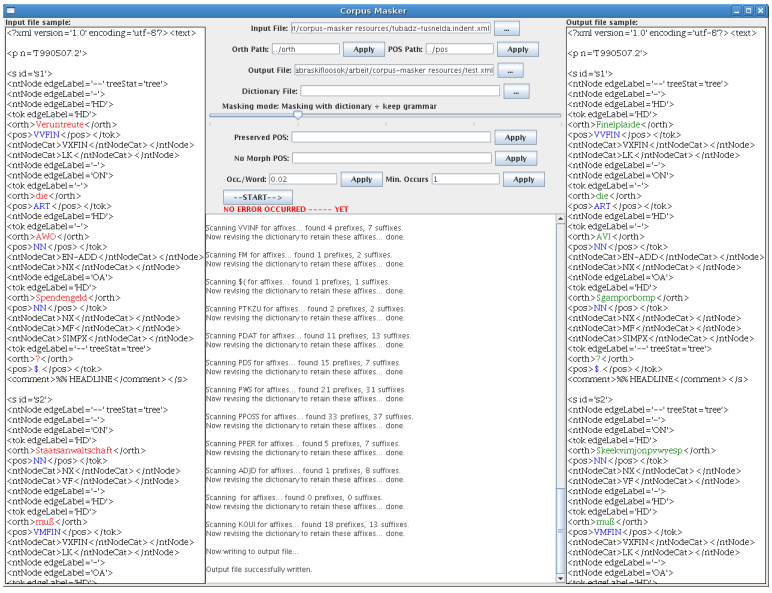

Figure 1: The graphical user interface of the CorpusMasker tool

CorpusMasker is a fully functional software tool for the parameterised masking of linguistic resources. The tool was implemented in Java and expects a (potentially very large) XML document instance containing the corpus as input; the XML data is read using SAX parsers. It is possible to specify the XML element(s) or attribute(s) that contain the actual words or tokens to be masked (in case of TüBa-D/Z, the `<orth>` element) as XPath expressions that refer to the child, descendant and attribute axes (e. g., `../orth`) via command line parameters or a graphical user interface (GUI), so that arbitrary corpus annotation schemes can be handled (see figure 1). Using the GUI’s preview function, the user can instantly observe the effects of parameter changes, so that the specific needs in terms of masking intensity and

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

preservation of useful structure can be met very efficiently without having to wait for a complete masking run to examine the results.

Next to the two abovementioned masking methods, the tool comprises a dictionary-based approach: first, CorpusMasker collects all word forms from all texts contained in the corpus to be masked. Then, every word is mapped onto a randomly generated string and replaced by that string. The length of the masked word can be retained, as well as information on the distribution as well as positioning of vowels and consonants in a specific word (vowels in the source word are mapped onto vowels in the random word, the same applies for consonants; variables can be set in order to specify a minimal randomisation distance). If a word is usually written with an initial lower case character and that word appears at a sentence-initial position with the first character being upper case, the same randomised word is used (e. g., "dort" → "kulp", "Dort" → "Kulp"). In addition, CorpusMasker performs an affix analysis. The algorithm examines certain affixes of words, masks the roots, but retains the affixes. With the affix analysis enabled, 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. Finally, the user can specify word classes that should not be masked, so that, for example, closed classes such as prepositions and determiners are left unchanged.

Linguistic corpora very often contain part-of-speech information so that the mapping process from genuine words to random strings of characters results in a list that acted as a key to unlock the masked version of the corpus, i. e., to reconstruct the STC. As a publication of this complete list would contradict the original purpose of the tool, we plan to provide only a reduced version of the file (see section 4). Although this reduced version does not contain the words from the STC proper, it can be thought of as a lexicon that maps the randomly generated words onto part-of-speech tags.

All features mentioned above can be activated, deactivated, and configured using CorpusMasker's command line options and arguments or its GUI, so that the person overseeing the operation is able to influence the masking process as much as possible. Furthermore, a randomly generated dictionary can be applied for masking a new corpus. As can be seen, the parameterised masking of linguistic corpora can be performed with several different degrees of retaining linguistic information, from the complete removal of the source text collection to a rather light but sufficient masking that keeps, e. g., closed word classes within the texts unchanged (see table 1; affixes are underlined).[6]

[6] A downloadable version of CorpusMasker will be available on our web site under an Open Source license in the winter of 2007 (`http://www.sfb441.uni-tuebingen.de/c2/`).

Part-of-speech:	VVFIN	ART	NN	NN	
Original sentence:	Veruntreute	die	AWO	Spendengeld	?
	↓	↓	↓	↓	↓
Characters replaced with [xX9]:	Xxxxxxxxxx	xxx	XXX	Xxxxxxxxxx	?
	↓	↓	↓	↓	↓
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?"

The Masking Algorithm and its Implementation

CorpusMasker consists of two SAX Parsers: the first one (DictExtractor) extracts all the words from the XML elements or attributes specified on the command line, and assigns POS classes according to the elements specified by an XPath expression. The extracted tokens are sorted into a hash that maps POS classes onto lists of tokens and their replacement patterns. These patterns are created on the fly by applying random replacement of vowels and members of other sound/character classes with members of the same character class. The algorithm enforces changes, i. e., no character except for punctuation and one-letter-tokens may stay the same. The affix extraction works on the complete dictionary and is discussed in more detail below. The second parser (Replacer) uses the dictionary to convert the source document (the XML annotated corpus) into the output document (the masked corpus), by replacing the content of the `<orth>` elements with the patterns as defined in the dictionary. The conversion mappings are stored in a file that can be used by a demasking tool to reconstruct the original text of the STC. The format of the dictionary entries is: `[POS] [original] [replacement]`.

For affix extraction, CorpusMasker uses a brute force algorithm that operates on complete dictionaries for each POS class. First, the algorithm extracts all possible prefixes and suffixes from each word (e. g., a word such as "voran" would contain the candidate prefixes "v", "vo", "vor", and "vora" as well as the candidate suffixes "n", "an", "ran", and "oran") and stores them in a hash that keeps track of the number of occurrences of each candidate affix. In the second step, only the most frequent candidates are chosen as affixes. The threshold can be adjusted by two values: the relative occurrence rate, i. e., the number of occurrences of the candidate divided by the total number of tokens contained in the dictionary. This number may vary between 0 and 1. A value of 1 will cause the algorithm to accept only affixes that occur in every word with the POS, while a value of almost 0 will accept

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

anything that occurs at the beginning or ending of any word with the POS as an affix. According to our experiments, a value of 0.02 produces satisfying results for German (the affix has to occur in 2% of the words with the POS). Small POS classes with few members (primarily function words) will then be considered "affixes" so that they will not be changed during the masking process. If this is not a desired effect, the user may alternatively apply the second value, the minimum occurrence restriction (the minimum number of different words the affix must have been found in). A value of 10 results in the desired effect of forcing replacement of function words, and still ensuring robust affix recognition for larger word classes (where the relative occurrence rate will be the limiting factor). Finally, the selected affixes are applied to all tokens with a corresponding POS tag. For each dictionary entry the algorithm tests the presence of all prefixes and suffixes. If a prefix or suffix is detected, the respective affix will be restored in the replacement pattern.[7]

The algorithm has two shortcomings: first, it tends to interpret virtually every frequent word-initial and word-final string of letters as an affix. As a result, the first and the last letter of most replacement patterns are the same as in the original. This problem is rooted deeply within the algorithm, and the best way to get around it could either be not to allow single-letter affixes (certainly not feasible for all languages) or to ask the user to certify the affix status for every potential affix the algorithm has recognised. Second, compounds as well as inflectional forms are a genuine problem, as these cannot be either recognised or analysed in the masked version. Currently, "house" could be replaced by something like "yaima", but "houses" could be "zieles". One solution would be also to store the "stems" produced by subtracting affixes from words, so that these "stems" will always be replaced by identical patterns. However, new problems will arise if we attempt to extend the algorithm in such a way: the strict separation of POS classes could not be retained any longer because we would like the stem "gracious" to be replaced by the same pattern in words that belong to a different POS class (such as "graciousness"). Given those problems, we decided to stick to our very simple algorithm that will, nevertheless, preserve a surprisingly useful amount of morphological information in most cases.

4 Masked Corpora: What are They Good for?

Masked linguistic resources can be used in several different scenarios. Our original goal had been to give researchers and organisations interested in the TüBa-D/Z treebank the option of examining the annotation without

[7] This method ignores iterations of affixes (such as "*ver*-*un*-treute") as well as infixes (e. g., "zurück-*ge*-geben"). Stem variations such as "Haus" – "Häuser", "gehst" – "ging" – "gegangen" etc. pose an additional problem. As this approach is based on pattern matching that is insufficient for morphological analyses, we plan to integrate a morphological lexicon.

going through the potentially extensive process of ordering the corresponding CD ROM first; furthermore, some organisations may not be able to purchase the CD ROM due to financial restrictions. While these might be in the minority with regard to the rather inexpensive *taz* CD ROM, our approach might prove useful concerning the masking of resources that are based on a source text collection with a license that costs a four or five figure sum.

Sustainability platforms Section 2 described sustainability initiatives and the goal of building web-based platforms for the long-term archiving and distribution of linguistic resources. In order to enhance the security of the copyrighted data (in case of TüBa-D/Z, the STC), such a platform should be outfitted with the option of masking the downloadable corpus archive before a download. Should a lexicon that contains the mapping from German words to randomised strings ever find its way onto the internet, it could be used to reconstruct a few randomised versions of the resource only. Furthermore, a web-based and password protected dictionary lookup could be provided that enables researchers who downloaded the masked version to retrieve a small amount of randomised strings to German word translations. An amount of, for example, 50 lookups per month, is large enough to translate several sentences (e. g., for use in an educational course or in a publication) and small enough to prevent the complete resource from being reconstructed.[8] Another function could be full-text search in masked corpora: a user searches for *word* (performed behind the scenes in an unmasked corpus), all matches are extracted, the whole corpus is masked – except for *word* – and finally the matches are presented in a masked version, again, except for *word*.

Unlexicalised training A corpus distributed in a masked version can be used for all sorts of unlexicalised training. In the case of parsing, unlexicalised PCFGs trained on treebank annotations are demonstrated to be compatible with other unlexicalised parsers (Charniak, 1996). It is beyond doubt that lexical knowledge improves parsing performance but this does not necessarily require the lexicalisation of rules. In the case of TüBa-D/Z, e. g., a lot of relevant knowledge is encoded in the morphological layer and can be used even with a fully masked corpus. Klein and Manning (2003) show that an unlexicalised model can achieve a performance close to the state of the art for lexicalised models.[9] Furthermore, there are cross-linguistic differences and it is, e. g., argued that the effect of lexicalisation is negligible in the performance of German PCFGs (Arun and Keller, 2005, Dubey and Keller, 2003). Hinrichs et al. (2005) discuss experiments of memory-based

[8] The number of dictionary lookups and the corresponding period of time are dependent on the number of tokens in a resource. We are aware of the fact that this functionality can be, with regard to legal issues, considered a grey area at best (see section 6).

[9] Klein and Manning (2003) propagate the subcategorisation of closed-class categories such as PP[für] or PP[als]. This is possible in our scenario as well due to the option of keeping functional words unmasked (see section 3).

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

learning of anaphora resolution with respect to personal pronouns and reflexives. Their tool is trained on the annotation of TüBa-D/Z and does not take lexical information into account. Their features refer to morphological properties, parts-of-speech, syntactic boundaries and grammatical functions all of which are given in the annotation accompanying the masked source text. 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 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. For 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 (Levy, 2004, Steiner, 2006).

Teaching linguistics and computational linguistics The masking process masks the source text collection and generates a lexicon en passant (see section 3). As a consequence, the resulting resource contains an unnatural language that, in the case of TüBa-D/Z, acts like German syntaxwise. The lexicon of this language, however, is, for the most part, based on random strings of characters and maps these randomised strings onto part-of-speech tags. This very fact makes the masked treebank a valuable resource in the context of teaching linguistics, and computational linguistics. If students are forced to work with a language that has a known syntax and even a rudimentary morphology but lexical entries that bear no meaning whatsoever, they might be able to concentrate better on, for example, the tasks of developing grammar rules or improving parsing efficiency. This approach of blanking out the meaning of lexical items is compatible with Chomsky's notion of language as processing a set of symbols.[10]

Evaluating NLP software Another promising application scenario for CorpusMasker is the evaluation of language technology software. A lot of current NLP tools (taggers, parsers etc.) are based on statistical algorithms

[10] For centuries, typographers use a certain text fragment ("Lorem ipsum dolor sit amet. [...]") in order to evaluate new layouts and page designs without resorting to writing actual text or inserting multiple phrases such as "Content goes here, content goes here ...". The fragment of blind text gives the impression of being genuine text with 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).

that use n-gram language models extracted from annotated corpora and treebanks as training data. Employing a masked resource, it is possible to measure the actual influence syntax or tree annotations have concerning the precision and recall of these tools. For this purpose, the performance of a tool with regard to original corpora, as well as slightly and fully masked corpora can be compared by using these corpora as training and evaluation data in turn. This approach could result in substantial arguments for or against the use of treebanks as a resource for training NLP tools.

5 Related Work

The most directly related work in Computational and Corpus Linguistics concerns anonymisation, the removal of proper nouns and other identity-revealing phrases from texts in order to protect the privacy of the people mentioned (Corti et al., 2000, Rock, 2001). Poesio et al. (2006) describe an anaphora resolution-based anonymisation module that is able to replace both proper nouns such as "Grandpa Gaunting" as well as pronominal references to proper nouns. Medlock (2006) defines "anonymisation" as "the task of identifying and neutralising sensitive references" and presents a corpus of ca. 2,500 personal email messages, collected and anonymised using a machine learning technique. The anonymisation itself potentially involves the deletion of references to all kinds of names, addresses, titles, geographic and ethnic terms and so on. A second application area is concerned with the removal of cues that might reveal the identity of a text's author. A third area concerns the masking, or obfuscation of texts, as described in the present paper. We are not aware of other approaches to the masking of linguistic resources.[11]

6 Future Work and Concluding Remarks

In addition to publishing CorpusMasker and a masked version of TüBa-D/Z on our web site (see footnote 6), we plan to extend the functionality of the tool in several ways: in some cases the affix analysis fails and produces results that do not correspond to the linguistic properties of the processed words. We will deal with this problem by enabling the user who is overseeing the masking procedure to modify the list of affixes produced by the algorithm and to add further prefixes and suffixes for specific part-of-speech classes.

[11] In a message posted to Corpora-List on August 19th, 2006, Péter Halácsy suggested "sentence shuffling" as a method to distribute a copyrighted corpus under "fair use" conditions. The relevant part of the copyright notice Halácsy et al. apply to the Creative Commons-based license of the "Hunglish" corpus (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. If you are a copyright holder, and you consider the shuffled files infringing, please send email and we will remove the material in question from the corpus."

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

Moreover, we will experiment with the free definition of characters and their potential replacement characters, for example, to allow that labial sounds may be swapped freely, but plosive sounds may not. A feature such as this one implies that we need to adapt the algorithm for non-latin characters (for example, Cyrillic). Based on the two abovementioned functions we will integrate support for the representation of alphabets and affix lists in configuration files so that language-specific masking defaults can be provided.

We call our approach parameterised masking because the randomisation process itself can be influenced with regard to several parameters (see section 3). For example, one command line parameter can be used to specify word classes whose corresponding tokens should not be randomised. Typically, when closed word classes such as determiners and prepositions are kept intact, at least a minor part of the original meaning of a sentence can be guessed. Eventually, this will lead us to a very crucial question: what happens if we choose to mask only a very small number of words (for example, only proper nouns)? Do we have to mask every single word, or at least a certain percentage of words, in order to bypass the STC's licensing restrictions? When does the text that has been masked only minimally become the original text again, so that the license prohibited the distribution of the pseudo-masked linguistic resource?

References

Arun, A. and Keller, F. (2005): "Lexicalization in Crosslinguistic Probabilistic Parsing: The Case of French". In: *Proc. of the 43rd Annual Meeting of the Assoc. for Comp. Ling. (ACL'05).* Ann Arbor, Michigan, pp. 306–313.

Charniak, E. (1996): "Tree-bank Grammars". In: *Proc. of the Thirteenth National Conf. on Artificial Intelligence (AAAI-96).* MIT Press, pp. 1031–1036.

Corti, L.; Day, A. and Backhouse, G. (2000): "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).

Dipper, S.; Hinrichs, E.; Schmidt, T.; Wagner, A. and Witt, A. (2006): "Sustainability of Linguistic Resources". In: *Proc. of the LREC 2006 Satellite Workshop Merging and Layering Ling. Inf.*, edited by et al., E. Hinrichs. Genoa, Italy, pp. 48–54.

Dubey, A. and Keller, F. (2003): "Probabilistic Parsing for German using Sister-Head Dependencies". In: *Proc. of the 41st Annual Meeting of the Assoc. for Comp. Ling.* Sapporo, pp. 96–103.

Hinrichs, E.; Filippova, K. and Wunsch, H. (2005): "What Treebanks Can Do For You: Rule-based and Machine-learning Approaches to Anaphora Resolution in German". In: *Proc. of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, edited by et al., M. Civit. Barcelona, Spain, pp. 77–88.

Klein, D. and Manning, C. D. (2003): "Accurate Unlexicalized Parsing". In: *Proc. of the 41st Meeting of the Assoc. for Comp. Ling.* Sapporo.

Lehmberg, T.; Chiarcos, C.; Rehm, G. and Witt, A. (2007): "Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten". In: *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proc. of the*

Biennial GLDV Conf. 2007, edited by Rehm, G.; Witt, A. and Lemnitzer, L., Tübingen: Gunter Narr, pp. 93–102.

Lehmberg, T.; Rehm, G.; Witt, A. and Zimmermann, F. (2008): "Preserving Linguistic Resources: Licensing – Privacy Issues – Mashups". *Library Trends*. In print.

Levy, R. (2004): "The Statistical Properties of Coordinate Noun Phrases ". Presented at the Dept. of Ling., University of Colorado-Boulder, March 11, 2004.

Medlock, B. (2006): "An Introduction to NLP-based Textual Anonymisation". In: *Proc. of the Fifth Int. Conf. on Lang. Resources and Eval. (LREC 2006)*, edited by et al., N. Calzolari. Genoa, Italy, pp. 1051–1056.

Newman, P. (2007): "Copyright Essentials for Linguists". *Language Documentation & Conservation* 1 (1): pp. 28–43.

Piez, W. (2004): "Way Beyond Powerpoint: XML-driven SVG for Presentations". In: *Proc. of XML 2004*, edited by Wood, L. IDEA, Washington.

Poesio, M.; Kabadjov, M. A.; Goux, P.; Kruschwitz, U.; Bishop, E. and Corti, L. (2006): "An Anaphora Resolution-Based Anonymization Module". In: *Proc. of the Fifth Int. Conf. on Lang. Resources and Eval. (LREC 2006)*, edited by et al., N. Calzolari. Genoa, Italy, pp. 1191–1193.

Rehm, G.; Eckart, R. and Chiarcos, C. (2007a): "An OWL- and XQuery-Based Mechanism for the Retrieval of Linguistic Patterns from XML-Corpora". In: *Int. Conf. Recent Advances in NLP (RANLP 2007)*, edited by et al., G. Angelova. Borovets, Bulgaria, pp. 510–514.

Rehm, G.; Witt, A.; Zinsmeister, H. and Dellert, J. (2007b): "Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections". In: *Digital Humanities 2007*. ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign, pp. 166–170.

Rock, F. (2001): "Policy and Practice in the Anonymisation of Linguistic Data". *Int. Journal of Corpus Ling.* 6 (1): pp. 1–26.

Schmidt, T.; Chiarcos, C.; Lehmberg, T.; Rehm, G.; Witt, A. and Hinrichs, E. (2006): "Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources". In: *Proc. of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*. East Lansing, Michigan.

Steiner, I. (2006): "Coordinate Structures: On the Relationship between Parsing Preferences and Corpus Frequencies". In: *Pre-Proc. of the Int. Conf. on Linguistic Evidence 2006*. Tübingen.

Telljohann, H.; Hinrichs, E. and Kübler, S. (2004): "The TüBa-D/Z Treebank – Annotating German with a Context-Free Backbone". In: *Proc. of the Fourth Int. Conf. on Lang. Resources and Eval. (LREC 2004)*. Lisbon, Portugal.

Telljohann, H.; Hinrichs, E.; Kübler, S. and Zinsmeister, H. (2006): "Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)". Technical Report, Universität Tübingen.

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

Trilsbeek, Paul and Wittenburg, Peter (2006): "Archiving Challenges". In: *Essentials of Language Documentation*, edited by Gippert, Jost; Himmelmann, Nikolaus P. and Mosel, Ulrike, Berlin, New York: Mouton de Gruyter, pp. 311–335.

Varga, D.; Halácsy, P.; Kornai, A.; Nagy, V.; Németh, L. and Trón, V. (2005): "Parallel Corpora for Medium Density Languages". In: *Int. Conf. on Recent Advances in NLP (RANLP 2005)*, edited by et al., G. Angelova. Borovets, Bulgaria, pp. 590–596.

Zimmermann, F. and Lehmberg, T. (2007): "Language Corpora – Copyright – Data Protection: The Legal Point of View". In: *Digital Humanities 2007*. ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign, pp. 162–164.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476