MANCHESTER 1824

The University of Manchester

The GerManC Corpus

31/03/2012

GerManC project team

Martin Durrell
Paul Bennett
Silke Scheible
Richard J. Whitt

GerManC

School of Languages, Linguistics and Cultures
The University of Manchester
Oxford Road
Manchester
M13 9PL

1. The GerManC corpus: Introduction

The aim of the GerManC corpus project was to compile a representative historical corpus of written German for the years 1650-1800. A central initial objective was to provide a basis for comparative studies of the development of the grammar and vocabulary of English and German and the way in which they were standardized, and the structure and design of the GerManC corpus was intended to parallel that of similar historical linguistic corpora of English, notably the *ARCHER* corpus[1] and the *Helsinki corpus of English texts*[2]. But consistent attention was paid to maintain compatibility with corpus projects in Germany covering earlier historical stages of German, initially within the framework of the DDD project (*Deutsch Diachron Digital*), and latterly with the various parts of the *Historisches Referenzkorpus des Deutschen*, which are currently being compiled at various centres in Germany[3]. The idea for the project goes back to an initiative by Anita Auer (now at the University of Utrecht), who completed a doctorate in Manchester on in 2005. Dr Auer's work drew attention to the lack of corpus-based data for German during this period compared to English; she suggested undertaking the compilation of such a corpus for German and completed some preparatory work on it.

Following the model of the ARCHER corpus and given the aim of representativeness, the GerManC corpus consists of text samples of about 2000 words from eight genres: drama, newspapers, sermons and personal letters (to represent orally oriented registers) and narrative prose (fiction or non-fiction), scholarly (i.e. humanities), scientific and legal texts (to represent more print-oriented registers). In order to facilitate tracing historical developments, the whole period was divided into fifty year sections (in this case 1650-1700, 1700-1750 and 1750-1800), and an equal number of texts from each genre was selected for each of these sub-periods. This periodization follows the model established for the Bonn corpus of Early New High German[4] which is currently being updated as part of the *Historisches Referenzkorpus des Deutschen*. Given the areal diversity of German during this period, the corpus also aimed for representativeness in respect of region, and to this end broad regional divisions were adopted for the GerManC corpus, i.e. North German, West Central German, East Central German, South-West German (including Switzerland) and South-East German (including Austria), taking an equal number of texts for each genre and sub-period from these five regions, as shown in the following table:

[1] See http://www.llc.manchester.ac.uk/research/projects/archer/
[2] See http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/
[3] See http://www.deutschdiachrondigital.de/
[4] See http://www.korpora.org/Fnhd/

Periods	Regions	Genres
1650-1700 1700-1750 1750-1800	North West Central East Central West Upper East Upper	Drama Humanities Legal texts Letters Narrative prose Newspapers Scientific texts Sermons

The complete corpus thus consists of 360 samples, comprising approximately 800,000 words. Appendix 1 contains a lists of the files in the corpus with full documentation in an Excel spreadsheet. In relation to this it should be noted that some newspapers, especially from the first subperiod 1650-1700, were short and did not contain 2000 words. In such cases further text was taken from another newspaper (or, in one case, more than one other) newspaper from the same city within the period in order to complete an appropriate sample of at least 2000 words. A pilot project involving a single genre (newspapers) was successfully undertaken with support from the Economic and Social Research Council (ESRC) from 1 March 2006 to 31 March 2007 (grant no. RES-000-22-1609), with Professor Martin Durrell as Principal Investigator, Dr Paul Bennett as Co-Investigator, and Dr Astrid Ensslin as Research Associate[5]. Dr Ensslin left at the end of the pilot project to take up a post as Lecturer in New Media at Bangor University.

Following a positive evaluation of the pilot, the full project was approved for support by the Economic and Social Research Council jointly with the Arts and Humanities Research Council in early 2008 (grant no. RES-062-23-1118). Work on it started in September 2008 with Dr Silke Scheible (now at the University of Stuttgart) and Dr Richard J. Whitt (now at the University of Strathclyde) joining Professor Martin Durrell and Dr Paul Bennett as Research Associates, and the project was completed in August 2011[6].

The complete corpus is now available through the Economic and Social Data Service archive[7], the Oxford Text Archive[8], and the project website[9]. At present seven of the

[5] Durrell, Martin, Astrid Ennslin and Paul Bennett (2007), 'GerManC. A historical corpus of German 1650-1800'. *Sprache und Datenverarbeitung* 31, pp. 71-80.

[6] Durrell, Martin, Silke Scheible, Richard J. Whitt and Paul Bennett(2011), 'Investigating diachronic grammatical variation in Early Modern German. Evidence from the GerManC corpus'. In: Marek Konopka, Jacqueline Kubczak, Christian Mair,Frantisek Sticha, and Ulrich H. Waßner (eds.), *Grammatik und Korpora 2009*, (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, 1). Tübingen: Gunter Narr Verlag, pp. 539-549.

[7] http://www.esds.ac.uk/

[8] http://ota.ahds.ac.uk/

planned eight genres are generally accessible through these archives, the genre 'letters' will be released following the solution of outstanding access problems and potential copyright issues, but interested scholars may apply to Professor Martin Durrell for access (martin.durrell@manchester.ac.uk).

The corpus is available in a number of forms, as detailed in the following sections, and each is contained in a separate folder.

2. Raw corpus (RAW)

The **'raw'** corpus consists of a **simple transcription** in a text file, subsequently checked by **double-keying**. This is in principle identical to the original texts, with no annotation. Only characters that are available from the English keyboard are used, with other characters given in Unicode (a detailed list of Unicode characters used is given below in section 3.8). The minimum of words per text is 1900, the maximum 2300 (although this was sometimes exceeded in the genre 'drama' to allow for repeated names of speakers and stage directions). Passages in other languages, as well as tables and long lists are kept to a minimum in text selection, as are passages in verse (with the exception of verse texts in the genre 'drama').

Transcription conventions:

- The great majority of the transcribed texts are printed in ***Fraktur*** ('black letter') and these were transcribed simply by substituting the corresponding letter in Roman font, with the exceptions noted below in the remainder of this section. Where *antiqua* (Roman font) was used in a text otherwise printed in black letter, this is indicated by an annotation in the TEI file, but not in the RAW corpus.
- **Umlaut** variants are given in Unicode as written, e.g. 'ü' for *ü*, and 'uͤ' for *u* superscripted with *e*.
- **Ligatures** in Latin dipthongs are represented as such, using Unicode character map for keystrokes (e.g. æ for *æ*)
- ***'sz'/'ß'/'ss'*** are retained, with *ß* given in Unicode as ß
- ***'u'* and *'v'*** are retained in both upper and lower case, as in, for example, *vmb* or *vnd* as they occur in older texts. Similarly, lower case *'i'* and *'j'* are kept (e.g. *'ietzt'*), but upper case *'J'*, which is used without distinction in all texts in Fraktur, is given as *'I'* or *'J'* corresponding to the modern norm.
- **Capitals** are given as they occur, e.g. in titles, e.g. *'Sonntagischer MERCURIUS'*, in the first word of a text or paragraph, e.g. *'OB wolln...'*, or in '*GOtt*'.
- **Nasal bars are** given in Unicode, e.g. n̄ or ē (for *ē*). In the latter case no attempt is made to guess whether '*n'* or *'m'* may have been intended.
- **Hyphenation** is adapted to contemporary standards, and '-' is used throughout (rather than '=').

[9] http://www.llc.manchester.ac.uk/research/projects/germanc/

- **Virgulas** are given as such, e.g. *'Smirna/ den 7. May.'*
- **Indentations** are not represented.
- **Damaged or missing passages** were only marked as such in the TEI files. Gaps were left without any mark-up in the raw corpus.
- **Unclear/ambiguous word forms** were transcribed as they were understood (we deduced semantically and graphologically logical/feasible lexemes from context), but marked <unclear> in the TEI files.
- **Long *'s'*** is transcribed consistently as *'s'*, whether in Fraktur or antiqua. In practice all texts followed the norm that long *'s'* is used morpheme initially or internally, curly *'s'* in final position.
- Text in **running heads**, including page numbers (in headers/footers), was not transcribed.
- If first word of a new page is **repeated** at the bottom of previous page (as an aid to the printer), this was transcribed in the genre 'newspaper', but not in other genres.
- **Notes** of any kind were transcribed uncommented in plain text. **Footnotes** and **marginal notes** were inserted where appropriate to the text (without breaking sentences in the running text). Generally, all such text is included at the end of the paragraph in which it occurs - in square brackets in the raw corpus or with appropriate annotation, using the tag <note>, in the TEI version.

3. TEI-annotated corpus (TEI)

In principle all texts were annotated for structural elements using the tags prescribed within the framework of TEI lite P5 (cf.: www.tei-c.org), using the tools Exchanger XML Lite 3.2 Editor (in the pilot project for the 'newspaper' corpus), and oXygen 10.1 (or later) for the other genres which were compiled during the main project.

Given the experimental nature of the pilot corpus 'newspapers', this genre was initially annotated more intensively than the other genres. However, on the basis of the experience gained it was recognised that such intensive tagging was not time efficient, and simpler structural tagging was adopted for the other genres in the main project. Where particular tags or sets of tags were only used in the pilot 'newspaper' corpus, this is specifically indicated in the account below.

3.1 TEI Header

- **File description:** title statement with title of work and author (if known).
- **Publication statement:** place and date of publication.

- **Notes statement:** name of file, region, genre, period and details of selected extract.
- **Source description:** details of source of text

3.2 Structural mark-up

- **Paragraph divisions** and text sections or divisions as given in the original text (e.g. '`p`', '`div`', '`head`')
- **Font types** (i.e. changes to antiqua and/or italic ; e.g. '`hi rend="antiqua"`')
- In the genre '**Drama'**, any act or scene divisions, speakers ('`sp`', '`speaker`'), stage directions ('`stage`')
- Line groups and lines in verse ('`lg`', '`l`')
- **Abbreviations** are marked up ('`abbr`'), but not resolved, except in the genre 'newpapers'
- **Dates** are marked as such
- In the 'newspaper' corpus, **graphical features** (e.g. title boxes, borders, calligraphy, images): marked and described.

3.3 Passages in other languages and foreign words are marked as such, with an indication of the language concerned - but excluding words which can be convincingly identified as having been adopted as loans by this period. Greek and Hebrew words are not normally transcribed, but the presence of a word or phrase in Greek or Hebrew script is indicated by a gap marked for that language, e.g. `<gap xml:lang: heb>...</gap>`. The following abbreviations are used to tag words or phrases in the following languages:

ar	= Arabic	fra	= French	nl	= Dutch
cz	= Czech	gre	= Greek	rus	= Russian
deu	= German	heb	= Hebrew	pol	= Polish
eng	= English	it	= Italian	tur	= Turkish
esp	= Spanish	lat	= Latin	ung	= Hungarian

3.4 Notes of any kind were marked as such and integrated into the text itself with an attribute specifying the place in the original (e.g. '`foot`', '`margin-left`').

3.5 Quotations from other sources or texts (e.g. from the Bible) are indicated ('`bibl`'), as are references to other works. Quotes are marked as such only in direct speech (despite occasional quotation marks, reported speech is not marked as quotation).

3.6 Names (of places, people/gods, organizations) are given a unique individually tag in the 'newspaper' corpus only, using the abbreviations given in Appendix 2. This enables

queries in this part of the corpus for such names, e.g. for specific cities or historical personages such as ‘Constantinople’ or ‘King George II’.

3.7 Figures (e.g. newspaper logos/emblems, ornaments, lines) are tagged in the ‘newspaper’ corpus.

3.8 Special characters (and other diacritics, ligatures, symbols - including alchemical symbols) are entered in Unicode, i.e.:

ä	&#228;	**°**	&#176;	**ó**	&#243;	♃	&#9794;
Ä	&#196;	**Â**	&#194;	**Ō**	&#332;	☿	&#9791;
ö	&#246;	**â**	&#226;	**ō**	&#333;	♳	&#9795;
Ö	&#214;	**Å**	&#197;	**Ů**	&#366;	♄	&#9796;
ü	&#252;	**å**	&#229;	**ů**	&#367;	♅	&#9797;
Ü	&#220;	**Ā**	&#256;	**Ū**	&#362;	♆	&#9798;
ß	&#223;	**ā**	&#257;	**ū**	&#363;	♇	&#9799;
æ	&#230;	**ã**	&#227;	**ú**	&#250;	♈	&#9800;
œ	&#339;	**É**	&#201;	**Û**	&#219;	♉	&#9801;
Æ	&#198;	**Ē**	&#274;	**û**	&#251;	♊	&#9802;
é	&#233;	**ē**	&#275;	**Ù**	&#217;	♋	&#9803;
è	&#232;	**ë**	&#235;	**ù**	&#249;	♌	&#9804;
á	&#225;	**ê**	&#234;	**ÿ**	&#255;	♍	&#9805;
À	&#192;	**ę**	&#281;	**ȝ**	&#658;	♎	&#9806;
à	&#224;	**í**	&#237;	**1/4**	&#188;	♏	&#9807;
ò	&#242;	**ì**	&#236;	**1/2**	&#189;	♐	&#9808;
ç	&#231;	**ï**	&#239;	**3/4**	&#190;	♑	&#9809;
&	&#38;	**î**	&#238;	✚	&#10010;	♒	&#9810;
§	&#167;	**Ô**	&#212;	♀	&#9792;	♓	&#9811;
|	&#124;	**ô**	&#244;	♁	&#9793;	♃	&#9794;

Superscript ‘e’ as in ‘○ᵉ’ = superscripted letter followed by ͤ (e.g. aͤ - oͤ - uͤ)

Nasal bar = superscripted letter followed by ̄

4. Linguistically annotated corpus (LING)

In order to facilitate thorough linguistic investigation of the data, the annotated version of the GerManC corpus provides access to the following linguistic information:

1.) Word tokens;
2.) Sentence boundaries;
3.) Normalized spelling variants;
4.) Lemmas;

5.) POS tags;
6.) Morphological tags;
7.) Grammatical function & dependency information.

The linguistic annotation of this release of the GerManC corpus was implemented automatically without subsequent manual correction, and thus the annotations cannot be expected to be fully accurate. However, within the framework of the project automatic annotation tools were developed and applied specifically for Early Modern German in order to improve the annotation quality for the information listed under (1.) to (5.) above. In addition, for normalized spelling variants, lemmas, and POS tags the performance of the tools was assessed by means of a gold standard subcorpus of GerManC, which was developed as part of the project. This gold standard subcorpus aimed to be as representative of the main corpus as possible. In order that it should remain manageable in terms of annotation times and cost, the subcorpus was limited to only two of the three corpus variables, 'genre' and 'time', as it was established that the variable 'region' exhibited considerably less significant relevant variation. The gold standard therefore only included texts from the North German region, with one sample file per genre and time subperiod. The gold standard corpus thus contained some 58,000 tokens in total.

The following paragraphs provide an overview of the annotation format (4.1), and the annotation schemes and automatic tools which were used to produce the annotations (4.2-4.6).

4.1 Annotation format

Two versions are provided of the linguistic annotations in GerManC. The first is in GATE XML (**LING-GATE**), a well-documented stand-off annotation format, which combines both structural (TEI) annotations of GerManC and linguistic annotations in a single document.

GATE ("General Architecture for Text Engineering")[10] is an open source text engineering platform which is highly customizable and supports the following steps:

1.) **Loading and viewing** the corpus documents.
For written instructions, see:
- Introduction to GATE User Interface
  http://gate.ac.uk/sale/tao/splitch3.html#x6-300003
- Loading and viewing documents
  http://gate.ac.uk/sale/tao/splitch3.html#x6-320003.2
- Creating and viewing corpora
  http://gate.ac.uk/sale/tao/splitch3.html#x6-330003.3

[10] www.gate.ac.uk

Important: When loading a new GerManC document/populating a GATE corpus with GerManC documents, specify "encoding" as "UTF-8".

The GATE website also provides helpful videos: http://gate.ac.uk/demos/developer-videos/; cf. in particular Modules 1, 2, and 3.

2.) **Viewing and editing** annotations

- Viewing existing annotations

  http://gate.ac.uk/sale/tao/splitch3.html#x6-340003.4 (especially 3.4.1 and 3.4.2)

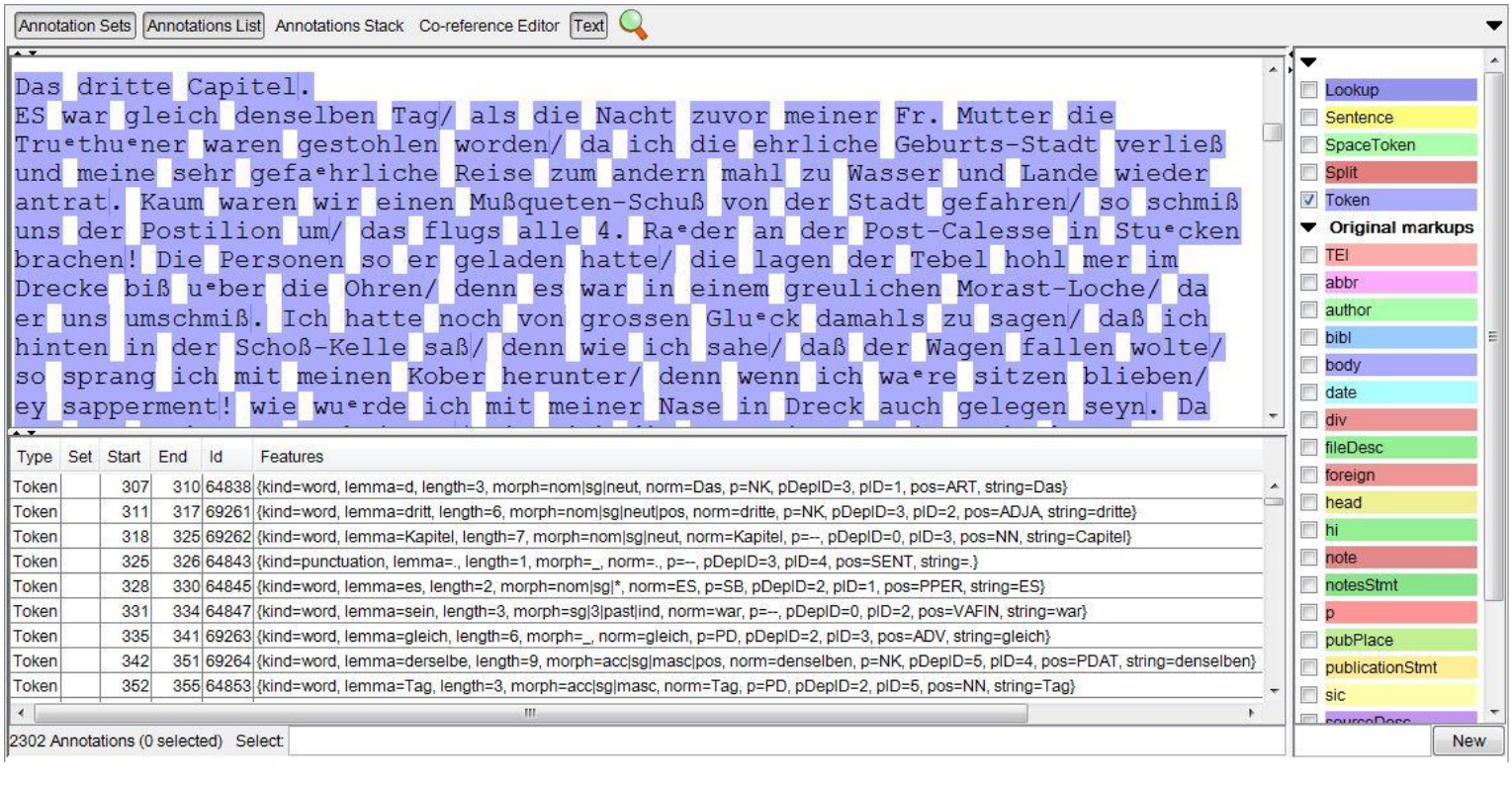
  The TEI annotations are stored in the Annotation Set 'Original markups'. The linguistic annotations are listed above the Original Markups set, as shown in the Screenshot below. To view the sentence mark-up, tick 'Sentence'. To view token-based mark-up, tick 'Token'.

  The following annotations are stored with tokens:

  - **'string'**: original word form
  - **'norm'**: normalized word form
  - **'lemma'**: base lemma (in its modern form)
  - **'pos'**: part-of-speech tag
  - **'morph'**: morphological information
  - **'p'**: syntactic category (parser output)
  - **'pID'**: word id in sentence (used by parser)
  - **'pDepID'**: dependency relation (parser output)

- Editing/adding annotations
  http://gate.ac.uk/sale/tao/splitch3.html#x6-390003.4.5

Cf. also video 'Module 6' at http://gate.ac.uk/demos/developer-videos/



3.) **Querying** the annotated corpus via the ANNIC Search GUI.
   - Written instructions:
     http://gate.ac.uk/sale/tao/splitch9.html#x13-2490009
   - Video tutorial:
     http://gate.ac.uk/demos/annic/annic.html

In addition to the GATE-based format, a text-based version of the linguistic annotations is provided in column format (**LING-COL**), where tokens (plus annotations) are printed in tab-separated rows, and empty lines indicate sentence breaks. For example:

pID	*string*	*norm*	*pos*	*lemma*	*morph*	*p*	*pDepID*
1	Das	Das	ART	d	nom|sg|neut	NK	3
2	dritte	dritte	ADJA	dritt	nom|sg|neut|pos	NK	3

4.2 Tokenization ('Token') and sentence boundaries ('Sentence')

Tokenization was carried out with a customized version of the GATE Unicode tokeniser and sentence splitter, which rely on finite state algorithms and the JAPE language. We added a number of rules to account for typographic variants typically found in EMG, such as certain ligatures (where two or more graphemes are joined as a single glyph, as in Æ) or combining letters such as a superscripted *e* in place of an umlaut (as in *o*^{e}). To achieve accurate tokenization, the appropriate unicode character classes were added to the tokenization rules. Further rules were added to treat hyphenated compound nouns as a single token, such as *Stadt-Kirche* ('town church'), *Slar-Affen* ('Cockaigne'), or *Aus-sicht* ('view'). These would also be treated as single tokens according to modern German orthographic norms (*Stadtkirche*, *Schlaraffen*, and *Aussicht*, respectively). Gazetteers were also added to identify full stops which indicate abbreviations. These were wrapped as single tokens, so that, for example |*kayserl.*| or |*Holl.*|, count as one token each.

The GerManC **sentence splitter** is based on an adaptation of the ANNIE sentence splitter for English in GATE, which defines sentence splitting rules based on punctuation symbols such as full stops "." (but, as explained above, excluding those that are part of an abbreviation), question marks "?", and exclamation marks "!". One of the most problematic issues concerning sentence boundary detection in Early Modern German is that punctuation is not standardized and varies considerably across texts. Conventional modern markers of sentence boundaries which are included in the ANNIE sentence splitting rules (such as full stops, exclamation marks, and question marks) sometimes do not occur at all in texts of this period. Instead, semi-colons, colons, and the virgule symbol "/" may have the function of marking both clause and sentence boundaries, and it is often difficult to decide which function was intended by the author. As this kind of variation is difficult to handle automatically, the sentence splitter takes the following general approach:

1.) Semi-colons and colons are added to the punctuation list indicating sentence splits.
2.) Virgule symbols are excluded.

This procedure aims to ensure that texts are divided up into useful chunks. However, it is important to be aware of the fact that semi-colons and colons may sometimes indicate clause boundaries, or even lists, rather than full sentences. The virgule symbol, on the other hand, is not included in the list of sentence splitters. Even though it is sometimes used to mark sentence boundaries, it is used in place of a modern comma in most cases, and thus not marking a sentence split.

NB. Sentence boundaries in the subcorpus 'newspapers' from the pilot project were marked manually.

4.3 Normalization ('norm')

Spelling variation is a well-known problem in the automatic processing of older language varieties. Non-standard spellings are particularly frequent in earlier texts in the GerManC corpus (ca. 35-40% of all tokens at the beginning of the early modern period, ca. 1650), while the proportion is lower in later texts (ca. 5-10% towards the end, ca. 1800).

Each token in the GerManC corpus was normalized to a modern form using Bryan Jurish's canonicalization tool, which is described in his doctoral thesis[11]. An evaluation of the output produced by the tool against the gold standard normalizations showed that it achieved high precision (96.9%) but relatively low recall (55.0%) on the task of identifying spelling variants, and 77.4% accuracy for the task of identifying the modern spelling of any detected variants. It should be noted, however, that the normalization guidelines adopted are relatively strict, so that, for example, the verb ending *-et* is normalized to the modern norm *–t* (*springet* → *springt*, 'jump'), although Jurish's guidelines would not require this.

4.4 Lemmatization ('lemma')

The lemmatization scheme implemented here aims to resolve each token in the corpus to a base lexeme in modern form, using the spelling prescribed by Duden[12] prior to the reforms from 1996 onwards. With obsolete words, the leading form in Grimm's *Deutsches Wörterbuch*[13] is taken. The guidelines for developing the gold standard are described in the introduction to Section 4. The lemma mark-up in GerManC was produced by a version of the TreeTagger[14] which was retrained on the basis of the POS and LEMMA information in our gold standard subcorpus, using the normalized word forms produced by Jurish's tool as input. In addition, the tagger's

[11] The thesis is available at: http://www.ling.uni-potsdam.de/~moocow/pubs/jurish2011diss.pdf
We are very grateful to Bryan Jurish (Potsdam) for making this tool available to us.
[12] http://www.duden.de/
[13] http://www.dwb.uni-trier.de/
[14] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

lexicon was extended with over 80,000 word forms from a modern corpus of German[15]. The retraining achieved an average accuracy of 86.83% on unseen data from all genres and time periods included in GerManC (using 'leave-one-out' cross-validation on the 24 gold standard files). To achieve maximum accuracy in tagging GerManC, the lexicon was further extended by manually adding POS and lemma information for the most frequent word forms in GerManC that had not yet been included in the lexicon (i.e. all word forms with a frequency greater than 8 in GerManC: 905 word forms altogether).

4.5 POS-tagging ('pos')

The GerManC POS tagging scheme is based on the STTS tagset for German[16], with a number of modifications to account for differences between modern and Early Modern German (EMG), and to facilitate more accurate searches. The STTS-EMG tagset thus contains a number of additional categories to account for special EMG constructions, such as various kinds of non-standard relative markers. The table below shows the STTS-EMG tag set with new categories highlighted in bold red. The new POS categories account for around 2.0% of all tokens in the Gold Standard subcorpus of GerManC.

The 'pos' annotations in GerManC were produced by the retrained version of the TreeTagger (mentioned above), which achieved 89.81% accuracy on our gold standard subcorpus (using 'leave-one-out' cross-validation). The model used for annotating the main corpus is likely to achieve better accuracy due to the extension of the tagger's lexicon with the most frequent word forms (as described above).

Tag	Description	Example
ADJA	attributive adjective (including participles used adjectivally)	das **große** Haus die **versunkene** Glocke
ADJD	predicate adjective; adjective used adverbially	der Vogel ist **blau** er fährt **schnell**
ADV	adverb (never used as attributive adjective)	sie kommt **bald**
APPR	preposition left hand part of double preposition	**auf** dem Tisch **an** der Straße entlang
APPRART	preposition with fused article	**am** Tag
APPO	postposition	meiner Meinung **nach**
APZR	right hand part of double preposition	an der Straße **entlang**
ART	article (definite or indefinite)	**die** Tante; **eine** Tante
CARD	cardinal number (words or figures); also declined	**zwei; 526; dreier**
FM	foreign words (actual part of speech in original language may be appended, e.g. FM-ADV/ FM-NN)	**semper fidem**
ITJ	interjection	**Ach**!
KON	co-ordinating conjunction	**oder** ich bezahle nicht

[15] We are grateful to Prof. Stefanie Dipper (Bochum) for providing us with this lexicon.
[16] http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

KOKOM	comparative conjunction or particle	er arbeitet **als** Straßenfeger so gut **wie** du
KOUI	preposition used to introduce infinitive clause	**um** den König zu töten
KOUS	subordinating conjunction	**weil** er sie gesehen hat
NA	adjective used as noun	der **Gesandte**
NE	names and other proper nouns	**Moskau**
NN	noun (but not adjectives used as nouns)	der **Abend**
PAV [PROAV]	pronominal adverb	sie spielt **damit**
PAVREL	pronominal adverb used as relative	die Puppe, **damit** sie spielt
PDAT	demonstrative determiner	**dieser** Mann war schlecht
PDS	demonstrative pronoun	**dieser** war schlecht
PIAT	indefinite determiner (whether occurring on its own or in conjunction with another determiner)	**einige** Wochen **viele solche** Bemerkungen
PIS	indefinite pronoun	sie hat **viele** gesehen
PPER	personal pronoun	**sie** liebt **mich**
PRF	reflexive pronoun	ich wasche **mich** sie wäscht **sich**
PPOSS	possessive pronoun	das ist **meins**
PPOSAT	possessive determiner	**mein** Buch das ist der **meine/meinige**
PRELAT	relative depending on a noun	der Mann, **dessen** Lied ich singe [...], **welchen** Begriff ich nicht verstehe
PRELS	relative pronoun (i.e. forms of *der or welcher*)	der Herr, **der** gerade kommt der Herr, **welcher** nun kommt
PTKA	particle with adjective or adverb	**am** besten, **zu** schnell, **aufs** herzlichste
PTKANT	answer particle	**ja, nein**
PTKNEG	negative particle	**nicht**
PTKREL	indeclinable relative particle	**so**
PTKVZ	separable prefix	sie kommt **an**
PTKZU	infinitive particle	**zu**
PWS	interrogative pronoun	**wer** kommt?
PWAT	interrogative determiner	**welche** Farbe?
PWAV	interrogative adverb	**wann** kommst du?
PWAVREL	interrogative adverb used as relative	der Zaun, **worüber** sie springt
PWREL	interrogative pronoun used as relative	etwas, **was** er sieht
TRUNC	truncated form of compound	**Vor-** und Nachteile
VAFIN	finite auxiliary verb	sie **ist** gekommen
VAIMP	imperative of auxiliary	**sei** still!
VAINF	infinitive of auxiliary	er wird es gesehen **haben**
VAPP	past participle of auxiliary	sie ist es **gewesen**
VMFIN	finite modal verb	sie **will** kommen
VMINF	infinitive of modal	er hat es sehen **müssen**
VMPP	past participle of auxiliary	sie hat es **gekonnt**
VVFIN	finite full verb	sie **ist** gekommen
VVIMP	imperative of full verb	**bleibt** da!
VVINF	infinitive of full verb	er wird es **sehen**
VVIZU	infinitive with incorporated *zu*	sie versprach **aufzuhören**
VVPP	past participle of full verb	sie ist **gekommen**

4.6 Morphological information ('morph') and parser output ('p', 'pDepID')

The morphological analysis and parse annotations included in GerManC were produced automatically by a tool for modern German, developed by Bernd Bohnet at IMS Stuttgart. Since there is currently no gold standard data for these annotation types figures cannot be provided for the accuracy of the tool. It produces high accuracy parses for modern German, but it is unlikely to have achieved this for Early Modern German texts, and users must be aware that it does not provide information suitable for quantitative analysis. The tool is described in:

> Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. The 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China[17].

The associated annotation schemes are based on the TIGER corpus[18] annotation guidelines for morphology and syntax:

Morphology:

> http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/annotation/tiger_scheme-morph.pdf

Syntax:

> http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/annotation/tiger_scheme-syntax.pdf

[17] This paper is available online at: http://www.aclweb.org/anthology/C/C10/C10-1011.pdf
[18] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/

5. Guidelines and notes for GerManC POS, LEMMA, and NORM annotations

These guidelines were developed and implemented in the production of the Gold Standard subcorpus.

5.1 General principles for annotation: features and categories

ABBREVIATIONS: Abbreviated words are assigned to the appropriate LEMMA in the unabbreviated form. The NORM has a form with an ending **if** this is clear from the text, e.g. *Herrn* for *Hrn*, but a form identical with the lemma if the text does not indicate an ending.

ADJECTIVE COMPARISON: the LEMMA for these is the positive form, i.e. *sicherer* and *sicherste* are assigned to the LEMMA *sicher*. However, emphatic superlatives such as *allergeringste* are assigned to a LEMMA *allergeringst*, since there is no positive form **allergering*. Superlative forms like *am schnellsten* are ADJD (and the *am* is PTKA)

ADJECTIVE INFLECTION: For all endings which do not correspond to modern usage, the modern norm is consistently used as the NORM, e.g. *himmlischen* in *die himmlische Scharen*, *kaltem* in *von kalten Wasser* and *häusliches* in *dein häuslich Leben*.

ADJECTIVES USED AS NOUNS have the tag NA. They are given capitals in all cases, and the lemma has the conventional ending *-e*. Thus, a form like *des Gesandten* has the NORM *Gesandten* and the LEMMA *Gesandte*; *folgendes* has the NORM *Folgendes* and the LEMMA *Folgende*. Comparative and superlative forms have a LEMMA deriving from the positive form, so that *etwas Günstigeres* has the NORM *Günstigeres* but the LEMMA *Günstigere*. NB that adjectives which refer back to a noun understood are still ADJA, e.g. *Ich habe nicht das rote Hemd gekauft, sondern das **blaue***. In modern usage these do not have capitals.

ADVERBS: The distinction between ADV and ADJD is difficult. The basic criterion is that any form which **can** be used as an attributive adjective should be ADJD, but there are some exceptions of lexicalized forms such as *früher* or *gleich* which are tagged as ADV. These are listed in the STTS guidelines.

CAPITALIZATION in NORM follows the modern norms, i.e. 'NORM' will have capitals after full stop, question mark or exclamation, but not after colon or semi-colon. Adjectives from place-names have small initials, e.g. *augsburgisch*.

COMPARATIVE PARTICLES *als* and *wie* are only tagged as KOKOM if they do not introduce a subordinate clause.

COMPOUND NOUNS written as a single word (or with a hyphen) have a NORM and a LEMMA corresponding to the modern norm, e.g. with no internal capitals or hyphens. Thus *Südseeinsel* for *SüdSeeInsel* or *Südsee-Insel* or any other variant on these.

Compounds written in the text as two (or more) words are tagged as individual words, so that *Süd See Insel* is tagged as three words. Although this is not wholly satisfactory, it is the only practical solution. A similar problem arises with verb prefixes which are written separately from the verb, e.g. *wahr nahm*, where in modern usage they would be written together. There is a case for tagging these as PTKVZ. This practice follows the principles in STTS (section 2.5).

CONJUNCTIONS consisting of two words. *als wenn* is treated by the tagger as KOKOM *als* KOUS *wenn*. This is extended as a model to similar cases, e.g. *ob* KOUS [...] *gleich* ADV, etc., as well as e.g. *ohne* APPR *daß* KOUS

CONTRACTIONS other than preposition + article, e.g. *hastu*, are re-tokenized as two words.

DATIVE *-E*: Masculine and neuter nouns are always given an endingless form in the dative singular, except where it is obligatory or usual in modern usage, e.g. *zu Hause*.

DEFINITE ARTICLE: Long forms of the definite article, e.g. *denen, derer/deren* are forms of the LEMMA *d* and have a NORM corresponding to modern usage, e.g. *den*, *der*.

DOUBLETS, by-forms, and the like are given a NORM and LEMMA corresponding to the most frequent modern equivalent. Thus *sonsten, dahero, jetzo/itzt, Friede, zumalen, weilen, darzu, worinnen, herfür* are taken as forms of *sonst, daher, jetzt, Friede, zumal, weil, dazu, worin, hervor*. In cases of doubt (e.g. whether *darinnen* a form of *darin* or *drinnen*?), the existence of a separate entry in Grimm's *Deutsches Wörterbuch* was taken as indicative.

FOREIGN WORDS: The criteria for distinguishing 'foreign' from 'native' words is explained in the STTS guidelines. In particular, all words (especially Latin) with non-German inflection are tagged FM. It is not necessary to give the part of speech in the original language. Also, all parts of foreign words and titles (but not personal or place names) are FM, e.g. *A* (FM) *Fish* (FM) *Called* (FM) *Wanda* (NE).

GENITIVE OF STRONG MASCULINE AND NEUTER NOUNS: Modern usage varies between the endings *-es* and *-s*. The form in the text is taken as the NORM.

GREETINGS like *Hallo* or *Adieu* are tagged as ITJ, as they are invariable and not part of any syntactic structure.

IMPERATIVES: The STTS guidelines only allow for the simple imperative forms of the second person, e.g., *Geh!* or *Geht!* but the verb-form in 'formal' *Gehen Sie!* is also tagged as VVIMP.

INDEFINITES: All instances of *alles*, *allem, alle* etc. are assigned to the LEMMA *all*. Similarly all forms of *dieser, dies, diese* are assigned to a single LEMMA *dies* (i.e. no distinction is made, as the STTS tagset appears to, between *dies* and

diese or *all* and *alle* as discrete lemmata). The same applies to all pronouns and determiners with a singular and plural form, which only have one basic LEMMA (normally the stem without any inflectional suffix). Those with only a plural, e.g. *beide* or *mehrere*, have a LEMMA with *-e*. Individual words are noted in the wordlist. Many indefinites can be used pronominally or as determiners; these are tagged PIS or PIAT respectively (we no longer use the STTS tag PIDAT), e.g. PIS: ***solche*** *sind selten*; PIAT: ***solche*** *Häuser* or *ein* ***solches*** *Haus*

INFINITIVE WITH *ZU: zu* + simple infinitive is often written together in this period. This was solved by re-tokenizing.

INTERJECTIONS: These include greetings (q.v.) and things like *Fa la la*.

LANGUAGES. Names of languages like *Englisch* in *auf Englisch* are tagged NN, but when used as adjectival nouns, e.g. *das Englische* they are NA.

LEMMA. The LEMMA is a base lexeme in modern form to which older forms found in our texts can be related, using *Duden* pre-reform spelling. With obsolete words, the leading form in Grimm's *Deutsches Wörterbuch* is taken.

MISPRINTS: Misprints are corrected in NORM and LEMMA in the annotation. The original text has the TEI mark-up <sic>.

MODAL VERBS: The *Ersatzinfinitiv* of modal verbs is tagged as VMINF when it is used to replace the past participle (e.g. *Er hat kommen* ***müssen***)

NAMES: All parts of names of places are tagged as NE, even foreign names, for example: *New* NE *York* NE. However, *von* in a personal name is APPR.

NAMES OF BUILDINGS, etc, In cases like *das Hohe Stift* the adjective can have the capital in the NORM, i.e. *Hohe*, but not in the LEMMA, which will be *hoch*.

NORM: The NORM is the form, inflected where appropriate, which the LEMMA would have in modern usage. In this way the NORM is the token, the LEMMA the type. Inflected nouns for the NORM have the prescribed modern form as given in *Duden*, so *Herzen* or *Kometen*, etc., and no account is taken of variation in modern German. Thus, the dative *-e*, which is already variable in this period, is normalized without the ending - except where it is required in modern usage, e.g. *zu Hause*, *unter Tage*. Thus, *Schwantze*, for example, has the NORM *Schwanz* and the LEMMA *Schwanz*.

NUMERALS: Declined forms of numerals (e.g. *zweier*) are tagged as CARD. Texts of this period often have a following full stop after cardinal numbers. This is ignored for the LEMMA and the NORM, which are given as in modern usage. Ordinal numbers are in principle ADJA. If they appear in the text in the form with a following full stop (e.g. **7.**) this can be used as the NORM and LEMMA, as can Roman numerals. Otherwise, with other forms of abbreviation, this is spelled out in line with what is in the text. Thus *6sten* has the NORM *sechsten*

and the LEMMA *sechst*. NB that *das Erste* (meaning 'the first thing') can be NA. Gender forms of *zwei* (*zwene, zwo*) are given the NORM *zwei*

OBSOLETE OR UNFAMILIAR WORDS were checked in Grimms *Deutsches Wörterbuch* and/or the DUDEN dictionary and the basic forms used there adopted (using the pre-reform spelling norms rules). Problematic instances were carefully considered and raised with other members of the team.

Archaic conjunctions, e.g. *wofern*, are tagged as KOUS, as is *wann* used in the sense of modern *wenn*. This was our practice for all words which are not used in accordance with their modern PoS function, e.g. *inwendig* as a preposition. Such words were consistently tagged according to their function in the period, ignoring in such cases the prescriptions relating to individual words in the STTS guidelines.

Care was taken with archaic forms, such as the past participle without *ge-* which is still quite frequent with some verbs (especially *kommen* or *funden*). These were tagged VVPP and given the NORM corresponding to modern usage, e.g. *gekommen*, *gefunden*. Similarly the now archaic *ward* (VAFIN) is given the NORM *wurde*.

PAST PARTICIPLES. It is tricky to differentiate between ADJD and VVPP. In general, particles in a participial phrase, e.g. *Der Kaiser, neulich aus Italien* ***gekommen,*** *war über seinen Sieg erfreut* are tagged VVPP. Participles with the auxiliaries *sein* and *werden* are also most frequently VVPP as part of passive constructions, but some, like *bekannt*, are lexicalized and are ADJD.

PERSONAL NAMES: Old case forms for personal names in principle have a modernized NORM, e.g. *Ladislas* (gen.) for *Ladislaen*, but *-en* is retained for feminines, e.g. *Beaten*.

POLITE PRONOUNS OF ADDRESS in the second or third person, e.g. *Er, Sie, Ihr* (used to single persons) are capitalized in all forms.

PRONOMINAL ADVERBS: The tagset uses PROAV for pronominal adverbs (*daran* and the like), whereas the manual has PAV. The latter form is preferred throughout. NB that *daher* is ADV, not PAV.

PROPER NAMES (in particular place names) are in principle lemmatized and normalized to their modern form, if one exists and can be identified. Otherwise the form in the text is used.

REFLEXIVE PRONOUN: *sich* is PRF and the tagger relates it to a LEMMA *er/es/sie*. This annotation is retained, since it is not always straightforward to identify which pronoun is involved.

SYNCOPE: adjectives and determiners with *-r-* are spelled out in full the NORM, e.g. *anderen* rather than *andren* or *andern* - with the exception of *euer*, which is almost never used in the full form (e.g. ***?euere*** *Bücher*) at any period. For

words in *-l-*, however, syncope is regular and retained, e.g. LEMMA *edel* NORM *edle*.

VERB ENDINGS in *-e*. The schwa is often subject to apocope, especially in dramas, (e.g. *ich komm*, *ich hätt*) but the NORM retains it. Similarly the ending *-et* is normalized to *–t* where this is the modern norm (e.g *sagt* not *saget*, but *wartet* is kept in that form).

VERB PREFIXES are tagged as such if they are written separately from the verb - often contrary to modern usage. *hin-* and *her-* compounds are not always prefixes; they can be adverbs, cf under these words in section 5.2.

WEAK FEMININE NOUNS could cause difficulties, since a form in the text like *der Gassen* at this period is potentially ambiguous between Genitive Plural (NORM *Gassen*, LEMMA *Gasse*) and Genitive/Dative Singular (NORM *Gasse*, LEMMA *Gasse*). The context may not always disambiguate satisfactorily, and the annotator's decision could be challenged.

5.2 General principles for annotation: specific words

This list is intended to complement the STTS guidelines document and provide information on particularly problematic words or relating to points where new tags were employed.

1. *Ahndung* is an archaic form of *Ahnung* and treated as a by-form of this
2. *alles* and all such forms are assigned to the LEMMA *all*.
3. *allergeringste* and similar emphatic superlatives are assigned to a LEMMA *allergeringst*, as there is no corresponding positive form **allergering*.
4. *allerlei* or *allerhand* are PIAT or PIS
5. *als* a decision whether KOUS or KOKOM is the appropriate tag is not always straightforward The STTS guidelines prescribe that KOKOM is **only** to be used **if** *als* does not introduce a subordinate clause, e.g. *größer **als** ich; er arbeitet **als** Taxifahrer*, but this is not always clear. In some older texts it has the meaning 'i.e.', in which case it is tagged KON.
6. *als wenn* is treated by the tagger as KOKOM *als* KOUS *wenn*. This is quite acceptable, and can also be applied to *als ob* and other conjunctions consisting of two words.
7. *am* with the superlative, e.g. ***am** besten* is PTKA
8. *Amen* is tagged ITJ
9. *ander*: *die **anderen** Menschen* is ADJA, *die **anderen*** is PIS; ***anders*** is ADV
10. *aufs* with the superlative, e.g. **aufs** Schlimmste is PTKA, followed by an adjectival noun which is tagged NA
11. *äußere* (and similar words, e.g. *obere*) has a LEMMA **with** *e*

12. *bange* often occurs in the form *bang*, but DUDEN gives *bange* as the chief variant.
13. *beide* has the LEMMA ***beide***
14. *bis* as a preposition is always APPR even when followed (as is usual) by another preposition. The tagger often tags it as KOUS even when it is clearly not a conjunction.
15. *bißchen* is PIS used on its own, PIAT with as following noun, cf *ein wenig*
16. *bloß* is ADV when used as a modal particle
17. *böse* (**with** *-e*) is given as the main form in Duden
18. *da* can be KOUS ('since', 'as') or ADV ('there')
19. *daher* and *dahin* should always be ADV, despite the ruling in the STTS guidelines.
20. *damit* is only PAV in the sense 'with it'; as the conjunction 'so that' it is KOUS.
21. *dann* in the sense of modern *denn* 'for', is KON (but the LEMMA is still *dann*); in the sense 'than' it is KOKOM.
22. *darzu* and similar forms of PAV with *-r-* before a consonant can be normalized to *dazu*.
23. *das* used as a pronoun, e.g. ***das****, was er sagt* or ***das*** *hast du davon* is PDS
24. *denn* in a text has this form as a lemma, even if it has sense of modern *dann*; in the sense 'than' it is KOKOM.
25. *dergleichen* (invariable) is PIS
26. *Dero* when used in a form of address (*Ihre Exzellenz kann* ***Dero*** *Pferd gehen lassen*) is an indeclinable PDAT, like *deren* (see below). It always has a capital, as a pronoun of address. Other instances may be a pronoun PPER.
27. *derselbe* and *derjenige* have this basic form for the LEMMA, i.e. not distinguished for gender. They are PDAT or PDS.
28. *dessen* (or *deren*) used as possessives, e.g. *Anna und deren Mann,* are PDAT.
29. *desto* is tagged KON
30. *dies*: all forms of *dieser, dies, diese* are assigned to a single LEMMA *dies*
31. *doch* can be KON or ADV depending on its function in the clause, cf the STTS Guidelines.
32. Contracted forms like *dran*, *drin* are treated as doublets of *daran* and *darin*, but *drinnen* is distinct (analogous to *draußen*)
33. *dräuen* is treated by Grimm as a by-form of *drohen*.
34. *ehender* 'previously' is listed separately by Grimm.
35. *eigentlich* used as a modal particle is ADV
36. *einander* is PRF
37. *einer* as an indefinite pronoun, e.g. ***einer*** *ging* or *die* ***einen***, is PIS; *der* ***eine*** *Arm* is ADJA.

38. *einerlei* is PIS; or PIAT if qualifying a noun.
39. *einige* has the LEMMA *einig*.
40. *entweder … oder*: both parts are KON
41. In *auf Erden*, the noun has the regular norm *Erde*, depite the fact that the archaic ending is still retained in this idiom.
42. *erst*: *die **ersten** Menschen* is ADJA, *die **ersten*** is PIS, *das Erste* is NA.
43. *erstere* 'the former' is PIS
44. *etc.* is ADV and the NORM and LEMMA have this form.
45. *etliche* has the LEMMA *etlich*
46. *etwan* is a distinct LEMMA from *etwa*, cf Grimm.
47. *etwas*: *etwas Brot*, *etwas Besseres* is PIAT; *etwas langsamer* is ADV; *sie hat etwas gesehen* is PIS
48. *ferner* in the sense of 'further' is ADV
49. *folgende* can be PIS, e.g. *Sie behauptete **folgendes**, **das folgende** hat sie nicht beachtet*. The LEMMA has the form *folgende*
50. *fort* in e.g. *sie ist fort* is ADV (cf guidelines)
51. *Frieden* (**with** *-n*) is the modern norm, cf Duden
52. *früher* is ADV in the meaning 'formerly'
53. *für* in text always has this form as a LEMMA, even when used in the sense of modern *vor*
54. *fürnehmlich* is a distinct LEMMA from *vornehmlich*
55. *ganz* is ADV in the meaning 'quite'
56. *gemelt* in the sense 'abovementioned' is a LEMMA in its own right, not a form of *melden*.
57. *gern* and *gerne* are by-forms of the same LEMMA. Duden gives *gern* as the chief form. The comparative forms *lieb* and *am liebsten* are treated as forms of this LEMMA.
58. *gleich* is ADV in the meaning 'immediately': and APPR when used with a noun, e.g. *gleich meinem Vater*
59. *Gott* is to be tagged as NE when referring to the Christian God. It can be ITJ.
60. *gut* is ADV in contexts like *gut 10 Kilo*
61. *haben* is **always** tagged as VA-, even when used as a full verb.
62. *hin-* and *her-* forms are usually verb prefixes, i.e. PTKVZ, but they can be ADV **if** they are not part of an established separable verb. The STTS guidelines are not helpful here.
63. *hingegen* is ADV
64. *hinkünftig* can only be used adverbially and is thus always ADV
65. *hintere*, *innere*, etc. have a LEMMA **with** *-e*

66. *indes*: Grimm and Duden treat this as a less frequent by-form of *indessen*
67. *inskünftig(e)*: According to Grimm the base form is *inskünftige*, and it is only ADV
68. *insonderheit* can be ADV
69. *ja* can be PTKANT as an answer word; ADV when used as a modal particle; ITJ when simply as interjection to start a dialogue turn.
70. *je* KOUS [...] *desto* KON
71. *jüngst* 'recently' is ADV
72. *mit kurzem* (i.e. 'in short') or *vor kurzem*, etc. The STTS guidelines are unsure about such forms and say they are 'ADJA, *möglicherweise* ADV'. They are treated as ADV as this is the more plausible tag.
73. *lange* as ADV (= 'for a long time') is distinct from the ADJA/ADJD *lang* (and always has a final *-e*). Similarly *längst*.
74. *lauter*, as in *lauter Verrückte* is PIAT according to the guidelines, and distinct from *laut*.
75. *letztere* 'the latter' is PIS
76. *lieber* and *am liebsten* are comparative and superlative forms of the ADV *gern*.
77. *maßen* can be a conjunction at this period, meaning 'since', 'because'.
78. *mehr* and *meist* are comparative and superlative forms of the ADV *viel*. NB that *mehr* can be used adjectivally or adverbially in this period, e.g. *ein Mehres*, which is tagged ADV.
79. *mehrer-* has the LEMMA ***mehrere***
80. *der mein(ig)e* and the like are tagged PPOSAT when a noun follows, otherwise PPOSS
81. *mit* can be ADV, e.g. in *das gehört* ***mit*** *zu deinen Aufgaben* cf STTS Guidelines
82. *nachdem* is normally KOUS, but some older texts use the form written together for the combination of preposition and article (= modern *nach dem*). This is tagged APPRART. If the conjunction is split, it is tagged *nach* APPR *dem* KOUS
83. *nacher* is a form of the preposition *nach*, cf Grimm.
84. *näheste* has the NORM *nächst* from the LEMMA *nahe*
85. *nämlich* is ADV when it is used to mean 'because'
86. *natürlich* is ADV in the meaning 'of course'
87. *nichts* is normally PIS unless followed by NA, e.g. *nichts Gutes*, when it is PIAT
88. *niemals* and *niemalen* are separate LEMMAs, according to Grimm.
89. *noch* is KON when part of the combination *weder .... noch*
90. *ob ... gleich* is KOUS ... ADV
91. *ob* is always KOUS, even when it does not introduce a finite clause. It is distinct from the preposition (APPR) *ob* meaning 'because of'.
92. *obere* has a LEMMA **with** *e*

93. *ohne* APPR… *daß* KOUS (cf STTS Guidelines)
94. *recht haben* and *mir ist etwas recht*. As the pre-1997 spelling norms are used, *recht* has a small initial letter for NORM and LEMMA in these constructions.
95. *sein* is always tagged as VA-, even when used as a full verb.
96. *selb* e.g. in *im* ***selben*** *Monat* is PDAT
97. *selber, selbst, selbsten* referring back to a person (e.g. *ich selbst*) are PRF
98. *selbig-* is tagged as PDS or PDAT, like *derselbige*
99. *sicher* is ADV in the meaning 'certainly', e.g. in isolation
100. *so* can be a relative particle PTKREL, or a conjunction meaning 'if', i.e. KOUS.
101. *sonsten* is a by-form of *sonst* and can have that NORM and LEMMA
102. *teils … teils* are both ADV
103. *trübe* (**with** *-e*) is given by Duden as the more usual form
104. *über* and *unter* can be ADV, e.g. *über fünf Meter hoch*
105. *um so* is *um* APPR *so* ADV, but *umso* is KON, cf STTS Guidelines
106. *vermutlich* is ADV in the meaning 'presumably'
107. *viel*: ***viele*** *Bücher* or *er trinkt* ***viel*** *Wein* is PIAT; *er trinkt* ***viel*** is PIS; *wir haben* ***viel*** *gelacht* or ***viel*** *mehr als du* is ADV.
108. *von* in personal names, e.g. *Hubert* ***von*** *Kerlwitz* is APPR; *von* APPR *….her* APZR
109. *vor* in text always has this form as a LEMMA, even when used in the sense of modern *für*
110. *vorbei* is always PTKVZ, cf STTS Guidelines
111. *wahrlich* is ADV in the meaning 'truly'
112. *wann* in the sense of modern *wenn* 'if', is KOUS (but the LEMMA is still *wann*)
113. ***was*** *macht er* is PWS; *das,* ***was*** *ich mache* is PWREL NB: *was* has the LEMMA *etwas* when it is a short form of this, but with the NORM *was*.
114. *weder … noch*: both parts are KON
115. *weit* is ADV when modifying a comparative, e.g. *weit mehr als…*
116. *weiter* is ADV when it means 'further', but ADJD as comparative of *weit* 'wide'
117. *weitläuftig* is a by-form of *weitläufig*, cf Grimm
118. *wenig*: (*ein*) *wenig Geld* is PIAT, (*ein*) *wenig* is PIS, *wir haben wenig gelacht* is ADV
119. *wenn* (see *wann*)
120. *werden* is always tagged as VA-, even when used as a full verb.
121. *wie*: ***Wie*** *geht es dir?* is PWAV; *er will wissen,* ***wie*** *es ihr geht* is PWAV; *er erklärte,* ***wie*** *man es macht* is KOUS (although these are difficult to distinguish); *so schnell* ***wie*** *möglich* or *Einrichtungen* ***wie*** *Krankenhäuser* is KOKOM. KOKOM is **only** used **if** *wie* does **not** introduce a subordinate clause.

122. *wo*, *worüber* (i.e. *wo*+preposition) are PWAV **unless** used as a relative pronoun, when they are PWAVREL. *wobei* can be KOUS if used in its modern sense.
123. *woselbst* is a conjunction, i.e. KOUS
124. *Zeitlauft* is a distinct, if archaic word which only occurs in the plural form *Zeitläufte*. Grimm gives it in the form *Zeitlauft*, and this is probably best LEMMA
125. *ziemlich* is ADV in the meaning 'fairly'
126. *zukünftig* is ADV in the meaning 'in future'
127. *zum* with the superlative, e.g. *zum Besten* is PTKA