VOICE

Vienna Oxford International Corpus of English

Part-of-Speech Tagging and Lemmatization Manual

(1st revised version May 2014)

Contents

1 Introductory remarks[1]

Part-of-speech tagging (POS tagging), i.e. the assignment of word class categories to tokens in a corpus, has become a standard feature in corpus annotation. The obvious advantage of POS tagging for corpus users is that it enhances the searchability of a corpus, since it provides additional information about the (corpus) data which corpus users would otherwise have to laboriously work out for themselves.

While there are, of course, large-scale corpora of L1 data whose spoken components are also part-of-speech tagged (e.g. BNC, COCA), there are to date no fully POS-tagged corpora of spoken L2 data, let alone English as a lingua franca (ELF) data. POS-tagging VOICE was in many aspects different from traditional POS tagging; in the absence of suitable models to refer to, the part-of-speech tagging of VOICE was a challenging and time-consuming process, carried out between 2009 and 2012. In fact, the tagging process itself raised a number of questions, e.g. about the (im)possibility of clear-cut categorization of intrinsically variable language. Given this particular condition, it might be helpful to make a few introductory remarks about the implications for POS tagging within an ELF corpus framework:

Firstly, it is important to stress that POS tagging is, just like any form of annotation, necessarily only an approximate process: the information it provides is always to some degree a function of subjective interpretation. Language use is of its very nature intrinsically variable, and could not function as a means of communication otherwise, so the idea that it can be definitively categorized into distinct parts of speech must always be understood as to some degree a convenient descriptive fiction, albeit a useful and widespread one which linguists and language professionals make use of in the description and teaching of language, and which is recorded/codified in grammar books. However, when criteria for categorization are specified, how far particular instances of actual use meet these criteria is often problematic. There are times when linguistic forms and/or their co-textual connections give sufficient evidence for their grammatical categories to be assigned with some degree of confidence. But there are also many cases when the evidence is inconclusive.

Secondly, most POS tagging has to date been carried out on corpora of native speaker data, predominantly written, and it is this kind of data that tagging procedures have been developed to deal with. Thus, written L1 (English) data can be annotated by direct reference to established grammars and tagging procedures. Where problematic cases occur, decisions to assign one part of speech tag or another to a linguistic form can be informed by familiarity with what 'normally' occurs in native speaker usage. There can be no such appeal to 'normality' in the POS tagging of VOICE. The data are quite different, consisting of spontaneous and, to a large extent, highly interactive speech events capturing the spoken usage of English not as a native language but as a lingua franca, where the usual conventions of seeming L1-normality do not apply. The speakers in VOICE interact with each other by exploiting the resources of English in varied and nonconventional ways. Not surprisingly, the occurrence of many non-canonical forms in the ELF data poses somewhat of a challenge when trying to apply conventionally codified word class categories in the process of POS

[1] In this 1st revised version of the *VOICE Part-of-Speech Tagging and Lemmatization Manual*, dated May 2014, a number of errata contained in the original version were corrected. The authors would like to thank Nora Dorn and Claudio Schekulin for their much appreciated help with the revised version of this document.

tagging. Essentially, relying fully on existing English language tagging practices for VOICE would have constituted an attempt to apply a system of annotation to data it was not designed to account for.

This naturally places a particular premium on interpretation. In POS tagging VOICE, we were thus faced with making decisions about how to tag forms which did not meet the criteria for conventionally codified categorization – forms, for example, that were morphologically marked as nouns but did not syntactically function as such, e.g. the word 'sticker' in *I cannot **sticker** this*. But since they needed to be tagged in one way or the other, operational decisions needed to be taken, e.g. by categorising items which did not easily fit into a particular category, and thus ruling out other, perhaps equally legitimate, options. Such operational decisions are, to some degree, always bound to be arbitrary, but unavoidable in the process of POS tagging, which necessarily involves the static categorization of what is intrinsically variable language use. In the tagging of spoken ELF data as recorded in VOICE, this becomes especially apparent.

And herein lies the particular significance of POS tagging for a corpus like VOICE. Our aim was to modify existing tagging categories and procedures to arrive at a POS-annotated version of VOICE so as to make it more user-friendly. In the process of producing VOICE POS, the problems of applying conventional categories to spoken ELF data made us aware of the essential features of any natural language use that are made particularly evident in the use of English as a lingua franca – features that are perhaps sometimes not perceived when dealing with native speaker usage because they are so familiar.

In short, the process of adding POS tags to VOICE was in many cases far from straightforward: the variation in the data and lack of a fully relevant precedent to follow posed many challenges. These, we dealt with as best we could in the tagging scheme described in this manual. All these challenges, and the way they were met, are further discussed in Osimk-Teasdale (in prep.) and Radeka (in prep.).

Vienna, January 2013, Barbara Seidlhofer, Ruth Osimk-Teasdale, Michael Radeka

2 Technical remarks on the tagging of VOICE

2.1 Tokenization

Tokenization is the process of dividing up a text into separate grammatical segments, e.g. into individual tokens, or sentential elements. This segmenting into tokens is a pre-processing step for other annotations such as part-of-speech tagging and lemmatization. The tokenization in VOICE included an extraction of 'pure' text (without mark-up), as well as pauses and laughter from VOICE XML, and the splitting of the text into individual parts, i.e. into tokens.

Verb contractions and the genitive *'s* are split into their individual components. Each component then receives a separate POS tag, as demonstrated below.

Text	Tokenization into individual parts	Part-of-speech Tagging
it's	it + 's	it_PP 's_VBS, it_PP 's_VHS, let_VV 's_PP, president_NN 's_POS
you're	you + 're	you_PP 're_VBP
we'll	we + 'll	we_PP 'll_MD
gonna	gon + na	gon_VVG na_TO
won't	wo + n't	wo_MD n't_RB
student's	student + 's	student_NN 's_POS
students'	students + '	students_NNS '_POS

2.2 Part-of-speech tagging

This section explains the technical procedures for part-of-speech tagging VOICE. The first operational decision anyone annotating a corpus with parts of speech has to take is the choice of a tagger and a tagging methodology. As VOICE is the first corpus of English as a lingua franca to be annotated with part-of-speech tags, there were no directly comparable models or easily transferrable tagging methodology for our data. In order for the workload to be feasible, we could not, however, start from scratch in designing our own guidelines and taggers either. We therefore had to rely on the adaptation of existing tagging models for the POS tagging of VOICE.

2.2.1 Adaptation of tagset

We chose to work with the tagset from the *Part of Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991) because this tagset uses rather coarse-grained categories, which reduce the number of possible ambiguities. This suited the needs for tagging the varied, spoken nature of the ELF data in VOICE. As expected, an analysis of the tagger's output on larger portions of our data revealed that we shared problems faced by previous approaches to POS tagging (e.g. ICE-GB, BNC) which applied and adapted taggers originally designed to deal with written language to typical features of spoken language. Such challenges are, amongst others, disfluencies, repetitions, re-starts, discourse markers and pauses. Additionally, tagger and tagset could not account for a number of features characteristic of our data, e.g. the input of multilingual speakers, including code-switches, non-canonical forms, and non-canonical form-function relationships. In order to account for these features we extended both the tagset of the Penn Treebank (cf. 3.3 VOICE Tagset, tags marked in green) and the tagging formats (cf. 3.2. Tagging formats in VOICE).

2.2.2 Creating the VOICE Tagging Lexicon

A POS tagger's lexicon lists word forms together with their possible tags according to a predefined tagset. Most of the available POS taggers are trained solely on the Wall Street Journal (Marcus, Marcinkiewicz & Santorini 1993). For the lexicon used for tagging VOICE, we drew on a number of different sources in order to build a lexicon which was as comprehensive as possible. We did not use the lexica which come with different taggers but chose to build our own in order to best cover the spoken, interactive and linguaculturally diverse features of VOICE data. In the VOICE lexicon, we included written, as well as spoken, Penn-Treebank corpora (Wall Street Journal, Brown, Switchboard, ATIS-3), large lexical databases like Wordnet (1995, Fellbaum 1998) and CELEX (Baayen, Piepenbrock & Gulikers 1995) and a number of word lists from our reference dictionary, which was the Oxford Advanced Learner's Dictionary, Edition 7 (OALD7, Hornby et al. 2007). Additionally, we consulted and added other resources to the VOICE tagging lexicon, e.g. those provided by the Natural Language Toolkit (NLTK) and GATE, as well as tokens from the frequency lists of the BNC (Kilgarriff 2006) and COCA (Davies n.d.). Finally, all of these resources were made compatible with the VOICE tagset, i.e. either by assigning tags where there were none, or adjusting the tag categories to those in the VOICE tagset. In a last step, the lexicon was completed with manually entered information: a few thousand words in the VOICE data remaining without an entry in the lexicon were added, e.g. many VOICE specifics, as the input of multilingual speakers, as well as markers of spoken language, and proper nouns. Where needed, additional tags were added to any tokens in the lexicon in order to account for non-canonical functions in VOICE. For example, for *partly* in the sequence *a partly answer*, we allowed for the tag JJ, in addition to RB. This last step was a dynamic process throughout the duration of the tagging of VOICE, in which new functions were added as we encountered them.

2.2.3 Choice of taggers and tagging procedure

After an initial trial run of a tagger on our spoken ELF data, in which we tested a number of utterances from VOICE on TreeTagger (Schmid 1994), achieving an accuracy of only 83-86% (Osimk-Teasdale 2013), we also tested other taggers, e.g. the Stanford tagger (Toutanova et al. 2003) and LTAG (Shen, Satta & Joshi 2007), achieving similar results; we furthermore attempted to improve accuracy by implementing hybrid systems (Brill & Wu 1998; van Halteren, Daelemans & Zavrel 2001; Wu, Ngai & Carpuat 2004; Radeka 2009) and by applying domain adaptation techniques (Daume III, Kumar & Saha 2010; Radeka in prep.). As expected, all of these achieved much lower accuracy rates on VOICE data (Radeka in prep.) than on the formal written native language they were trained on. In order to achieve higher tagging accuracies, we adapted both tagset and tagging strategy to the needs of our data. In order to make an informed decision for the part-of-speech tagging of VOICE, we took account of the various tagging procedures used by state-of-the-art taggers, i.e. Decision Trees (Schmid 1994), Maximum Entropy (Ratnaparkhi 1996; Toutanova et al. 2003), Support Vector

Machines (SVM) (Giménez & Marquez 2003), Transformation-based Learning (TBL) (Brill 1995), Memory-based Learning (Daelemans et al. 1996), Conditional-Random-Fields (Lafferty, McCallum & Pereira 2001), as well as Markov Models (Brants 2000). All these achieve similar accuracies on different types of data. In order to establish procedures relevant to our data, we chose to use a combination of three different types of taggers within a stacked-TBL framework, similar to N-fold Templated Piped Correction (Wu, Ngai & Carpuat 2004). This procedure of using three different taggers meant that the advantages of each individual tagger could be capitalized on. TBL is a system which corrects errors by formulating rules from the tagger output in comparison with a correctly annotated version. The advantage of these rules (in contrast to other tagging systems) is that they are transparent and therefore interpretable for a human annotator. Thus, TBL can also be used to detect regularities in forms of rules. These rules can be used as diagnosis tools for individual tagging systems, as well as data analysis tools, in that they can highlight interesting aspects of the data. TBL helps to analyse where and why tagging errors occur, and, as a consequence, allows the annotator to create manually crafted rules, which help to improve tagging accuracy (cf. Volk & Schneider 1998). Another important criterion for a suitable tagger for our data was a high degree of flexibility. TBL met this need as it allowed us to keep the lexicon separate from the disambiguation, and hence to adapt the lexicon, as well as the tagset and tagging format of the Penn Treebank, for our specific purposes.

The three taggers which we used in combination for the stacked TBL framework were TreeTagger (a Decision Tree tagger), Stanford Tagger (a bi-directional Maximum Entropy tagger), and LTAG (a bi-directional perceptron-like tagger). These taggers were first applied to 5 000, later to 10 000 words of data and subsequently compared to the manual annotation of these data. These 10 000 words of data had been manually tagged, the tagging results compared and, in diverging cases, discussed, by 2 project researchers. The manually and the automatically annotated data were then used as training sources for TBL. What TBL did was to generate rules based on the differences of these annotations in order to increase the tagging accuracy. In a second step, the remaining, not manually annotated, part of VOICE was tagged with the three taggers and the rules generated in the first step were applied to the taggers' output. Now the most frequently applied tags by the three taggers were chosen for each token, in a procedure commonly referred to as Voting (Brill & Wu 1998). From this, tag probabilities for each token in VOICE were calculated and added to the VOICE lexicon. This now complete lexicon, based on the estimated tag probabilities occurring in VOICE, was then used to initiate a parallel TBL procedure (Radeka 2009) in which replacement (Brill 1995) and reduction rules (Lager 2001) are learned. Through the analysis of these rules we gained systematic insights into the strengths and weaknesses of the taggers' output, which in turn highlighted which aspects a) could be fixed by rule modification or manual rule creation, and b) which needed to be annotated manually. In fact, this resulted in rather large parts of the corpus being annotated and disambiguated manually, e.g. discourse markers, multi-word items, long-distance dependencies, words standing alone or at the beginning or end of utterances. This manual annotation in combination with the stacked TBL procedure helped to further increase tagging reliability. We intend to carry out a more detailed analysis of the final tagging accuracy of the VOICE corpus in future research.

2.3 Lemmatization

Lemmatization is the process of grouping together word forms that belong to the same inflectional paradigm and assigning to each paradigm its corresponding uninflected form, called a 'lemma'. This form of annotation is related to the process of part-of-speech tagging, as the latter is a prerequisite for lemmatization. Lemmatization is regarded as a highly useful type of corpus annotation because it provides additional search options.

The lemmatization of VOICE was carried out by an especially designed lemmatizer which was implemented in Python. This lemmatizer accessed the VOICE Lexicon in its final, completed version and retrieved the appropriate lemma from there. The lemmatizer did this by applying a number of manually implemented rules, which were related to the information contained in the POS tags (e.g. for regular verb forms: if tag for word X is VV (verb, base form), then lemma equals token), or, sometimes, also to the morphological information contained in the tokens (e.g. for regular adjective forms: if tag for word X is JJR (adjective, comparative), then lemma equals token without the suffix -*er*).

3 The VOICE part-of-speech tagging guidelines

3.1 Guiding principles

The main goal of POS tagging VOICE, apart from gaining insights into the data as such, was to develop a tagging procedure and a scheme as appropriate as possible to the ways in which English as a lingua franca is used in the interactions recorded and transcribed in the corpus. This meant it was important that tagging should be compatible with the variable character of English as a lingua franca, and a perspective in which the speakers are primarily viewed as language users in their own right, rather than 'language learners' (Osimk-Teasdale 2013).

The main principles that guided the process of part-of-speech tagging VOICE were the following:

a) We gave an **ELF perspective** priority at all times (see e.g. Seidlhofer 2011). This means that we generally asked ourselves first and foremost which tags it made most sense to assign to specific positions from an ELF point of view. This had priority over the second step, in which we asked how these decisions could be carried out practically, i.e. how to proceed in accordance with good practice. Our ELF guiding principle sometimes resulted in decisions that posed complex challenges to carrying out the manual and automatic part-of-speech annotation and the technical implementation of tagging decisions.

b) For tag categorization and technical implementation, we relied on **established procedures** as far as possible. This means that whenever established tagging procedures used for other corpora or available descriptions of English(es) were compatible with our ELF perspective, we adopted these. The *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991) served as a starting point for the VOICE Tagset. These we modified and extended in order to account as appropriately as possible for the ELF data. This included using the format FORM(FUNCTION) (cf. 3.2 Tagging formats in VOICE), as well as modifying the tagset originally used for the Penn Treebank Project, including tags for various characteristics of spoken language and a large number of discourse markers. In addition, the 7th edn. of the *Oxford Advanced Learner's Dictionary* (OALD7) was used as external reference dictionary for tagging decisions[2].

[2] cf. Breiteneder et al. (2006: 179ff.) and Pitzl, Breiteneder & Klimpfinger (2008: 25) for the reasoning behind using OALD7 for VOICE.

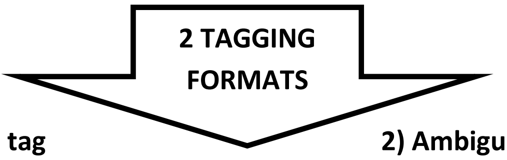
c) As a further guiding principle, we aimed at a tagging scheme which, while acknowledging the variable character of spoken ELF, would be **intuitively accessible** for researchers working with VOICE. We tried to achieve this by adhering to established points of reference, both **internal**, i.e. tagging guidelines in line with the already existing transcription conventions of VOICE (VOICE Project 2007b), as well as **external**, e.g. the *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991), the OALD7 (2005), and commonly acknowledged word class categories in general.

d) We tried to strike a **balance** between **inevitable interpretation**, i.e. **leaving options open** on the one hand, and avoiding **potentially excessive complexity** for corpus users, on the other. The former we did by making frequent use of **ambiguous tags**, in the format, and by introducing the basic tagging format **FORM(FUNCTION)** for all tokens (cf. 3.2 Tagging formats in VOICE for an explanation). In order to reduce complexity, however, we only allowed a maximum of **two tags per token,** e.g. *so_DM/RB.*[3] As this guideline was sometimes difficult to implement, especially with verb and noun forms, we introduced **'generic'** verb and noun tags (V and N, respectively), as in the example *xxx remark_NN/**V** anyway***.** Here the token *remark* can be assigned the tag NN (noun), but the tags VVP (present tense verb) or VV (base form) are also possible, due to the lack of disambiguating co-text (*xxx* indicating unintelligible speech). This would, however, result in the assignment of more than two tags, and hence a generic verb tag (V) is given instead of two sub-specified verb tags (VVP,VV) (cf. also 3.3.2 The commented VOICE Tagset). In cases where more than two tags were possible for one token, and these could not be simplified with a generic verb or noun tag, the **tag 'unknown' (UNK)** was assigned. For example, in the sequence *xxx like xxx*, the tags VVP, VV, IN or DM would have been possible for 'like' (cf. 3.4.2 Tagging of individual elements), but no decision for one tag was possible due to the lack of disambiguating co-text, hence the tagging was *like_UNK*. Another way in which we tried to minimize the range of possibilities for interpretation was to use a **sequential tagging procedure** as consistently as possible. This means that wherever there was more than one way of interpreting a stretch of tokens, we assigned only those tags which conformed to a 'sequential', i.e. left to right, reading of the tokens. For example, in the stretch *this is the first slides,* the token *slides* was interpreted as noun with plural form and singular function (tagging: slides_NNS(NN)), due to the tokens *this is* preceding the token *slides*, indicating singular function. The (also plausible) reading of *is* having singular form and plural function instead, was not taken into account.

[3] NB: one "tag" always consists of a form-tag and a function-tag in VOICE POS Online (cf. 3.2. Tagging formats in VOICE).

3.2 Tagging formats in VOICE

Category	Format
Form and function tags **Format: TAG(TAG)**	For all tokens in the corpus, separate tags for paradigmatic form and syntagmatic function are assigned. The tag for form is indicated first, followed by a tag for function, given in brackets. Format: **FORM-tag(FUNCTION-tag)** There are **2 options** of this format: **OPTION 1:** form and function converge → identical form(function) tag is assigned, e.g. *a **house_NN(NN)*** **OPTION 2:** form and function do not converge → different tags for form and (function) are assigned, e.g. *two **house_NN(NNS)*** **NB:** The format FORM-tag(FUNCTION-tag) is relevant when working with VOICE POS Online, as users are able to search for form- and function-tags separately. The default search in VOICE POS Online always considers positions for both form- and function-tags. For example, the search *NNS* will yield all of the following results: • multicultural teams_**NNS**(**NNS**) • in one countries_**NNS**(NN) • three university_NN(**NNS**). For the sake of simplicity, for examples in this tagging manual only one tag will be indicated whenever it is implied that form-tag and function-tag converge. Hence, e.g. *the group* will be indicated as *the_DT group_NN*, not *the_DT(DT) group_NN(NN).* Both tags, i.e. FORM-tag(FUNCTION-tag), will be indicated in this manual only when form and function-tag do not converge. For any token, a **maximum number of two tags** is allowed, whereby a "tag" refers to a form and a function-component, as in e.g. *so_IN(IN)/RB(RB).*



2 TAGGING FORMATS

1) Non-ambiguous POS tag

(one word class category is assigned to a token)

Format: TAG	1 tag is assigned, e.g. *be_VB* → Format in VOICE POS Online: *be_VB(VB)*

2) Ambiguous POS tag

(two possible word class categories are assigned to a token)

Format: TAG/TAG	2 tags are assigned with tags in alphabetical order, separated by a slash e.g. *use a maltese word or **joke_NN/VVP** in maltese* → Format in VOICE POS Online: *joke_NN(NN)/VVP(VVP)*

3.3 VOICE Tagset

The list of part-of-speech tags used for annotating VOICE data with word class categories is called the VOICE Tagset. The VOICE Tagset consists of 69 different part of speech tags, which are largely based on the *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991). These were designed for written data and then extended for the SwitchBoard corpus (cf. Linguistic Data Consortium (LDC 1999). However, in the course of tagging VOICE, it became clear that both the tagset as well as the tagging format used for the Penn Treebank needed to be modified and extended for our kind of data. Some substantial changes were therefore made in order to make the tagset better suited for the character of spoken, interactive ELF data. All additions to the tagset used for the Penn Treebank, as well as all changes in a tag's categorization are marked in **green**. Section 3.3.1 provides an alphabetically sorted list with regard to categories. Examples and explanations for tags are provided in section 3.3.2.

3.3.1 The VOICE Tagset, sorted alphabetically according to categories

Category	Tag
Adjective	**JJ**
Adjective, comparative	**JJR**
Adjective, superlative	**JJS**
Adverb	**RB**
Adverb, comparative	**RBR**
Adverb, superlative	**RBS**
Anonymization	**NP**, additionally marked **a_** preceding the token
Breathing	**BR**
Cardinal Number	**CD**
Conjunction, coordinating	**CC**
Conjunction, subordinating	**IN**
Contracted 's	**DOS** = does **VBS** = is (=BES in Switchboard) **VHS** = has (=HVS in Switchboard) **POS** = possessive **PP** = personal pronoun ***us*** (PRP in Switchboard)
Determiner	**DT**
Discourse Marker	**DM** (single discourse markers) **FORM-tag(FUNCTION-tag:DM)** (multi-word discourse markers) (cf. 3.3.2)
Foreign Word (Non-English speech)	**FW,** additionally marked **f_** preceding the token
Formulaic Item	**FI**
Interjection	**UH**
Laughter	**LA**
List Item Marker	**LS**

Category	Tag
Noun, generic	**N**
Noun, plural	**NNS**
Noun, singular or mass	**NN**
Onomatopoeia	**ONO**
Partial Word	**XX**
Particle	**RP**
Pause	**PA**, annotated **_0, _1, _2,** … (numbers indicate pause duration)
Possessive Ending	**POS**
Predeterminer	**PDT**
Preposition	**IN**
Pronoun, personal	**PP**
Pronoun, possessive	**PP$**
Pronoun, relative	**PRE**
Pronunciation Variation and Coinages (PVC)	**FORM-tag:PVC(FUNCTION-tag),** additionally marked **p_** preceding the token (cf. 3.3.2)
Proper Noun, plural	**NPS**
Proper Noun, singular	**NP**
Spelt	**SP,** additionally marked **s_** preceding the token
Response Particle	**RE**
Symbol	**SYM**
there, existential	**EX**
to, infinitive use	**TO**
Unintelligible Speech	**UNI**
Unknown	**UNK**
Verb, base form	**VB** = verb be **VH** = verb have **VV** = all other verbs (all = VB in Penn Guidelines)
Verb, generic	**V**
Verb, gerund or present participle	**VBG** = verb be **VHG** = verb have **VVG** = all other verbs (all = VBG in Penn Guideline)
Verb, modal	**MD**

Category	Tag
Verb, past participle	**VBN** = verb be **VHN** = verb have **VVN** = all other verbs (all = VBN in Penn Guidelines)
Verb, past tense; includes the conditional form of the verb *be*	**VBD** = verb be **VHD** = verb have **VVD** = all other verbs (all = VBD in Penn Guidelines)
Verb, present, non-3rd person singular	**VBP** = verb be **VHP** = verb have **VVP** = all other verbs (all = VBP in Penn Guidelines)
Verb, present, third person singular	**VBZ** = verb be **VHZ** = verb have **VVZ** = all other verbs (all = VBZ in Penn Guidelines)
Wh-adverb	**WRB**
Wh-determiner	**WDT**
Wh-pronoun	**WP (VOICE:** only tagged **WP** when ***not*** used as a relative pronoun! → else **PRE tag)**

3.3.2 The commented VOICE Tagset, sorted alphabetically according to tags[4]

Tag	Explanation and examples
BR	**Breathing**, e.g. *hh, hhh, hhhh*
CC	**Coordinating conjunction**, e.g. *and, but, or*
CD	**Cardinal Number**, e.g. *one, twenty-eight* (**VOICE:** also including *zero*),
DM	**Discourse Marker**[5]**.** Discourse markers are words which have homonyms in other word class categories and can function as discourse markers. VOICE operates with a closed list. A distinction is made between SINGLE and MULTI-WORD Discourse Markers: **1) SINGLE WORD DISCOURSE MARKERS:** **Items:** *like, look, whatever, well, so, right* **Tag**: DM **2) MULTI-WORD DISCOURSE MARKERS** **Items**: *I mean, I see, mind you, you know, you see* **Tags:** Multi-word discourse markers are tagged with a conventional word class tag for FORM and the tag DM for (FUNCTION): *I_PP(DM) mean_VVP(DM)* *I_PP(DM) see_VVP(DM)* *mind_VVP(DM) you_PP(DM)* *you_PP(DM) know_VVP(DM)* *you_PP(DM) see_VVP(DM)*
DOS	for **contracted 's**, **DOS** = does, e.g. *Where**'s** she live?*
DT	**Determiner**, e.g. *a, the, that* Some items, such as *that*, are also tagged DT when occurring without a head noun (analogous to Santorini 1991: 8)
EX	***there***, existential
FI	**Formulaic Items**, includes all formulaic expressions which are in the closed list "VOICE Formulaic Expressions", e.g. greetings, farewells, thanks, apologies, wishes, miscellaneous expressions. (cf. 6.1. VOICE List of Formulaic Items)
FW	**Foreign word** (Non-English speech), e.g. *francais.* Additionally marked with the prefix f_ in VOICE POS XML and VOICE POS Online.
IN	**Preposition** or **subordinating conjunction**, e.g. *because, behind*
JJ	**Adjective**, e.g. *good*
JJR	**Adjective**, comparative, e.g. *better*
JJS	**Adjective**, superlative, e.g. *best*

[4] Since the *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (Santorini 1991) served as a starting point for the VOICE Tagset, we are also using the explanations and partly also the wording used there. Note that these guidelines differ in a number of ways from later revised versions which include some tag changes which had to be made for the bracketing procedure (Santorini 1995). All changes to the original Penn Tagging Guidelines and adaptations made for VOICE are marked in **green**.

[5] Please note: 'mind you' and 'you see' are included in this list of discourse markers and in the Appendix (6.3.2 Multi-word discourse markers, p. 31 below). Unfortunately, due to an oversight the tagging of these two discourse markers does not appear in the published versions of VOICE POS. However, lists of these two discourse markers as they occur in VOICE and with the appropriate tags can be requested by sending an e-mail to <voice@univie.ac.at>.

Tag	Explanation and examples
LA	**Laughter**, e.g. *@, @@, @@@*
LS	**List Item Marker**, e.g. *section d_LS*
MD	**Modal**, e.g. *can, could, might, may*
N	**Generic Noun Tag**, used instead of ambiguous noun tags, e.g. NN/NNS or NP/NPS, primarily in tagging where there is a difference in form and function, e.g. *to: (.) register (.) to our lectures? and (.) stuffs_VVZ(N)* (cf. also Generic Verb Tag)
NN	**Noun**, singular or mass, e.g. *house, water*
NNS	**Noun**, plural, e.g. *houses*
NP	**Proper Noun**, singular, e.g. *european union*
NPS	**Proper Noun**, plural, e.g. *the netherlands_NPS*
ONO	**Onomatopoeic noises**, all onomatopoeia are represented in IPA-signs and are additionally marked with the prefix o_, e.g. *o_kr_IPA* in VOICE POS XML and VOICE POS Online.
PA	**Pause**, annotated with an underscore, followed by a number indicating the length of the pause in seconds (0 referring to up to approximately 0.5 seconds), e.g. *_0, _1, _2, …*
PDT	**Predeterminer**, e.g. *all, both* when preceding a determiner
POS	**Possessive Ending**, e.g. for **contracted ’s**, **POS** = possessive, e.g. *maria theresia**'s_POS** eyes.*
PP	**contracted ’s**, **personal pronoun *us***, e.g. *yeah let's_PP do something*, **possessive** and **reflexive pronouns** without case distinction, e.g. *they_PP knew that*, *do it yourself_PP*
PVC	**Pronunciation Variations and Coinages**, all items annotated <pvc> </pvc> in the transcription process were assigned the FORM-tag **PVC** and a suitable part-of-speech tag for function. Tokens given the tag PVCs are additionally marked with the prefix p_ preceding, e.g. *p_associational_PVC(JJ)* in VOICE POS XML and VOICE POS Online.
PP	**Pronoun,** personal, e.g. *I, me, you, he*
PP$	**Pronoun,** possessive, e.g. *my, your, mine, yours*
PRE	**Pronoun,** relative. Closed list: *that, which, who, whom,* and *whose.*
RB	**Adverb**, most words that end in *-ly* as well as degree words, e.g. *quite, too, very*
RBR	**Adverb**, comparative. Refers to adverbs with the comparative ending *-er*, with a strictly comparative meaning, e.g. *they are better_RBR recognized*
RBS	**Adverb**, superlative, e.g. *the most_RBR important education*
RP	**Particle,** e.g. *set up_RP support*
RE	**Response particle**, e.g. positive and negative minimal feedback, e.g. *no, yes, yeah, okay, yep, nah* (cf. 6.3.3 Interjections)
SP	**Spelling out,** referring to spelt items which could not be categorized further (cf. 3.1.2.11 Spelling out), additionally marked with the prefix s_ in VOICE POS XML and VOICE POS Online.
SYM	**Symbol**, used for mathematical, scientific or technical symbols, e.g. *x_SYM axis*
TO	***to***, infinitive use, e.g. *to_TO introduce*

Tag	Explanation and examples
UH	**Interjections** **CATEGORIZATION FOR VOICE:** these are markers of spoken discourse (e.g. hesitation markers) which do not have homonyms in other word class categories (as opposed to tokens tagged DM), e.g. *er, erm, yipee, whoohoo, mm:, haeh; a:h, wow* (cf. 6.3.3. Interjections).
UNI	**Unintelligible speech**, e.g. *x, xx, xxx*
UNK	**Unknown**, used for words which are ambiguous between more than two word class categories, e.g. due to lack of co-text.
V	**Generic Verb Tag**, used instead of ambiguous verb forms e.g. VV/VVP VVD/VVN, primarily in tagging where there is a difference in form and function, e.g. *will be communicate_V(VVG) with* (cf. also Generic Noun Tag).

VB/VH/VV
(all = **VB** in Penn Guidelines)
	Verb, base form, subsumes imperatives, infinitives and subjunctives
	VB = verb be
	VH = verb have
	VV = all other verbs

VBD/VHD/VVD
(all = **VBD** in Penn Guidelines)
	Verb, past tense; includes the conditional form of the verb *to be*
	VBD = verb be
	VHD = verb have
	VVD = all other verbs

VBG/VHG/VVG
(all = **VBG** in Penn Guideline)
	Verb, gerund or present participle
	VBG = verb be
	VHG = verb have
	VVG = all other verbs

VBN/VHN/VVN
(all = **VBN** in Penn Guidelines)
	Verb, past participle
	VBN = verb be
	VHN = verb have
	VVN = all other verbs

VBP/VHP/VVP
(all = **VBP** in Penn Guidelines)
	Verb, present, non-3rd person singular
	VBP = verb be
	VHP = verb have
	VVP = all other verbs

VBS	for **contracted 's**, **VBS** = be, e.g. *Tom's an excellent teacher.*

VBZ/VHZ/VVZ
(all = **VBZ** in Penn Guidelines)
	Verb, present, 3rd person singular
	VBZ = verb be
	VHZ = verb have
	VVZ = all other verbs

VHS	for **contracted 's**, **VHS** = have, e.g. *She's bought a nice dress.*
WDT	**Wh-Determiner,** e.g. *what, which, whatever* **VOICE**: Not used for relative pronouns. Used for e.g. *what_WDT kind* (vs. *what_WP do you like),* also: *which, whichever, whatever.* Original Penn Treebank Guidelines: Wh-determiner e.g. *which*, and *that* when it is used as a relative pronoun.
WP	**Wh-pronoun**, e.g. *what, who, whom* **VOICE:** only tagged **WP** when ***not*** used as a relative pronoun, else tagged **PRE.**

Tag	Explanation and examples
WRB	**Wh-adverb**, e.g. *how, where, why, when* When used to introduce a relative or an interrogative clause.
XX	**Partial words,** e.g. *becau-* Corresponding to "Word fragments" in the VOICE Mark-up conventions (VOICE Project 2007c: 3), the absent part is indicated with a hyphen.

3.4 Further specifications on the tagging of VOICE

3.4.1 Tagging of individual categories

Category **Tagging practice**

3.4.1.1 ANONYMIZATION

All anonymized tokens are labelled with the prefix **a_[…]** and are tagged **NP**, e.g. a_[S1]_NP, a_ [org4]_NP, a_[place5]_NP

3.4.1.2 COLLECTIVE NOUNS

Collective nouns are tagged singular or plural depending on whether the following verb is singular or plural. This is in line with the Penn Treebank Guidelines (cf. Santorini 1991: 18).

For VOICE POS Tagging, we follow the rather broad definition of Carter & McCarthy (2006: 541), who state that a collective noun is "[a] type of noun referring to a group of people, animals or things", as well as the examples given in Carter & McCarthy (2006: 539) and Quirk (1997: 316f.). Not included in our definition of collective nouns are cases in which names of countries are used representatively for the population of a country, as in the example below. In these cases, the verb which follows is tagged with differing tags for FORM and (FUNCTION), e.g. *the rest of the country need_V(VVZ) it as well*

3.4.1.3 –ING CATEGORY

In dealing with ELF data, it was often extremely difficult to decide whether a word ending in *–ing* should be classified as verb, noun or adjective. Hence, it was decided that **all words** in VOICE ending in the morpheme *–ing* would be given a **uniform FORM-**tag, namely **VVG**, and a **(FUNCTION)** tag according to their **syntactic co-text**.

For this category, the **FORM-**tag **VVG** stands for any word ending in the morpheme *–ing* (potentially followed by a plural *-s* morpheme). For the **(FUNCTION)**-tag, we only differentiated between either **VVG** and **NN** or **NNS**: The tag **VVG** was given when the word functioned as a present participle, and also when used as a participial adjective. The function-tags **NN** or **NNS**, respectively, were given when the word functioned as a singular or plural noun.

Tagging examples:

1. Word ending in *–ing* functions as **verb** or a **participial adjective**:
   TAG=VVG(VVG), e.g. *swimming_VVG(VVG)* man.
2. Word ending in *–ing* functions as **noun:**
   TAG=VVG(NN), e.g. the real *meaning_VVG(NN)*

The only exception was made for words which end in *–ing* and are listed as **adjectives** in the **reference dictionary** (OALD7) and which we regard as lexicalised for the tagging of VOICE), and tagged with **JJ**, the tag for adjectives.

1. Word ending in *–ing* is an **adjective** in OALD7:
   TAG=JJ, e.g. *charming_JJ man*

–ING CATEGORY cont.

NB: **Compound nouns** with one of the parts ending in the morpheme *-ing* (e.g. *swimming pool*), were also tagged according to the procedure described above, e.g. deciding whether the individual parts of the compound functioned as nouns in their immediate co-text (tagging: *swimming_VVG(NN) pool_NN(NN)*). Thus, for these cases, we did not consult the reference dictionary OALD7, as we did for all other compounds listed in the VOICE List of Compound Nouns (Section 6.4). Hence, the compound noun combinations with one component ending in *-ing*, such as *swimming pool*, are not included in the VOICE List of Compound Nouns.

3.4.1.4 MULTI-WORD ITEMS

As **multi-word items** we understand sequences of tokens which, grammatically, seem to 'belong' together and thus, form a single unit.

All parts of a multi-word item are assigned **identical tags**, e.g. *per_RB se_RB; student_NN union_NN.* If the **head** of the multi-word item is marked plural, all parts are given a plural tag, e.g. *points_NNS of_NNS view_NNS, youth_NNS organizations_NNS*

There are **5 types of multi-word items**:

1. Compound Nouns (cf. 6.4 VOICE List of Compound Nouns)
2. Items in VOICE Multi-words (cf. 6.2 VOICE List of Multi-words)
3. Multi-word Discourse Marker (cf. 6.3.2 Multi-word discourse markers)
4. Multi-word Formulaic Items (cf. 6.1 VOICE List of Formulaic Items)
5. Proper Nouns and Names (cf. 3.4.1.7 PROPER NOUNS (NP,NPS) vs. COMMON NOUNS (NN,NNS))

3.4.1.5 PARTLY UNINTELLIGIBLE

Tokens of which parts are annotated as unintelligible (marked <un>x</un> in VOICE Online) are given the tag UNI. This means that the part that was intelligible to the transcriber is also assigned the tag UNI, e.g.
VOICE Online: ***super****<un>****x*** *</un>*
VOICE POS Online: ***superx****_UNI*

3.4.1.6 PRONOUNS

For the tagging of VOICE, a distinction is drawn between the following pronouns:

1. Personal pronouns (Tag: PP)
2. Possessive pronouns (Tag: PP$)
3. Relative pronouns (Tag: PRE)
4. Wh-pronouns (Tag: WP)

Other pronouns are not assigned an individual tag category but are subsumed under other part-of-speech categories. For example, **demonstrative pronouns** such as *this* in *it was very nice you let us do* ***this****,* are tagged DT, and **indefinite pronouns**, such as *someone* in ***someone*** *is waiting,* are tagged NN. The **reciprocal pronoun** *each other* is also tagged NN (cf. Biber 1999: 70f. for an overview of the different pronouns in English).

3.4.1.7 PROPER NOUNS (NP,NPS) vs. COMMON NOUNS (NN,NNS)

General guidelines:

1. The tag **NP** includes **Proper Nouns** (which belong to the category noun e.g. *America*) as well as **Proper Names** (i.e. a combination of a proper noun with other words as *United States of America*) if they refer to a single entity.
2. **External references**: In some cases we oriented ourselves towards our reference dictionary OALD7 and tagged as proper noun when it was capitalized there, e.g. with regard to **alcoholic drinks, festivals**. If necessary, other dictionaries and search engines were consulted.
3. **Multi-word tag for proper nouns:** For titles of films, books etc. we use a **multi-word tag**, i.e. every word is assigned the NP tag even if it is not a noun, e.g. *good_NP night_NP and_NP good_NP luck_NP.* This is an **open list** and is not included in the VOICE List of Multi-words (cf. 6.2).
4. **Compound nouns**: For proper and common compound nouns where the **head is plural**, the word preceding or following the head is tagged plural NNS or NPS respectively, e.g. *swimming_NNS pools_NNS, points_NNS of_NNS view_NNS.*
5. **Form(Function) tags:** For proper nouns and names we usually did *not* use OALD7 as an external reference for paradigmatic form and syntagmatic function, as OALD7 does not list the majority of proper nouns and names occurring in VOICE. Sometimes this would have also resulted in odd combinations of tags for form and function, e.g. *Goofy* is only listed as JJ in OALD7, but occurs in VOICE as the Disney character → we tagged *NP*, not *JJ(NP).*

Tag NP or **NPS** is used for:

- **Alcoholic drinks and brands** (if capitalized in OALD7), e.g. *desperados, beaujolais*
- **Car names** and **names for aeroplanes**, e.g. *audi, jumbolino, saabs*
- **Currencies**, e.g. *lek, rouble, lei, dinar*
- **Days of the week, months,** e.g. *tuesday, may*
- **Famous personalities**, **groups etc.**, e.g. *aristotle, the smiths*
- **Languages,** e.g. *finnish*
- **Names of people, places, institutions, companies, programmes**, e.g. *nato, erasmus*
- **Names of products,** e.g, *ajax*
- **Nationalities**: e.g. *dane*
- **Professional terminology**, such as terms for **mathematical concepts**, e.g. *cauchy fanapppi*
- **Recurrent festivities** and **public holidays**, e.g. *christmas, ramadan*
- **Religious and spiritual terms** e.g. *feng shui, catholicism*
- **Religious denominations**, e.g. *christian(s), muslim, jews, baha'is*
- **Titles of films, books, names of websites**, e.g. *guinness book of records, youtube*

Tag NN or **NNS** is used for:

- **Alcoholic drinks** (if not capitalized in OALD7), e.g. *tequila*
- **Chemical elements**, e.g. *lithium chloride*
- **Diseases**, e.g. *meningitis, flu*
- **Food and beverages**, e.g. *goulash, rooibush*
- **Ordinal numbers in dates,** *e.g. the first_NN of October*
- **Titles**, e.g. *doctor, missis* (unless occurring as part of a proper name, e.g. *queen_NP elizabeth_NP)*

3.4.1.8 RELATIVISERS

For relativisers, a distinction is made between **relative pronouns** (*that, which, who, whom, whose*), which are tagged **PRE** and **relative adverbs** (*how, where, why, when*), which are tagged **WRB**. In this, we follow the distinction between these two categories drawn by Biber et al. (1999: 608).

3.4.1.9 SPELLING OUT

Items which are spelt are tagged as if they were spelt out normally, e.g. *eu* = "European Union" = NP, *tv* = "television" = NN. This refers to English as well as non-English speech, e.g. *oebb* (Austrian federal railways, a company) is tagged NP, not FW. Items which are spelt are additionally marked with the prefix *s_* before the spelt item, e.g. *s_eu*

The sub-categorization is as follows:

- **CD** in place of a number, e.g. *if we have **s_x_CD** universities*
- **SYM** for mathematical symbols
- **LS** for list items
- **NN** or **NNS** for spelt items which stand for nouns or function as nouns, e.g. if they can be pluralized. The same holds true for spelt items which function as Proper Nouns or Names (Tag **NP** or **NPS**), Verbs (corresponding verb-tag, e.g. **VVP**), etc.
- **SP** in case of 'real spelling' or if the spelt item could not be identified further.

3.4.1.10 UNCERTAIN AND PARTLY UNCERTAIN SPEECH

Both uncertain and partly uncertain speech are not marked as such in VOICE POS, i.e. uncertain speech in VOICE Online marked with brackets '(…)' is treated as normal text in VOICE POS, and no longer indicated with brackets. These items are assigned a POS tag referring to the token without consideration of the brackets signalling uncertainty.
e.g. **Uncertain speech:**
VOICE Online: *yeah **(just about)***
VOICE POS Online: *yeah **just_RB about_IN***
e.g. **Partly uncertain speech:**
VOICE Online: *a variety of **instrument(s)***
VOICE POS Online: *a variety of **instruments**_NNS*

3.4.2 Tagging of individual elements

As with the tagset, we used the *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* as starting point for the tagging of individual tokens (cf. 1991: 23ff.). The items listed below are cases we encountered in our data which did not have a corresponding guideline in Santorini (1991), or cases in which we found the guideline was not suitable for our data. In these cases, other external references (e.g. OALD7, other corpora, dictionaries and grammars) were consulted in order to decide on a suitable tagging scheme.

Individual token(s)	Tagging practice
ain't	*ai_VVZ n't_ RB* *ai_VVP n't_RB* (*ai_VHZ n't_RB*; *ai_VHP n't_RB* would also be possible in theory but do not occur in VOICE)
altogether	*altogether_RB*
and so on	*and_CC so_RB on_RB*
and that	**Meaning 'and similar':** e.g. *faked diamonds and_CC that_DT*
as regards	*as_IN regards_VVZ*
as such	*as_IN such_DT*
as well as	*as_RB well_RB as_IN*
get rid of, rid	*get_VV,VVP rid_VVN of*
gonna	*gon_VVG na_TO*
got	1. If clearly identifiable as participle → tag VVN, e.g. *have got_VVN,* 2. If simple past, or no or too little co-text to identify as past participle → tag VVD, e.g. *she got_VVD*
gotta	*got_VVD ta_TO*, or *got_VVN ta_TO* (see criteria for distinguishing between VVD and VVN cf. *got*)
how come	*how_WRB come_VV*, e.g. *how come the austrian are perceived as hh as being drunk all the time*
like	**Verb Present**: e.g. *I like_VVP it* **Verb Base Form**: e.g. *you don't like_VV the people* **Conjunction, Preposition**: e.g. *something like_IN that; it looks like_IN a sauce, I'll do it like_IN this* **Discourse Maker**: e.g. *they put like_DM erm poison all around; I was like_DM*
never mind	*never_RB mind_VV*
no	**Adverb**: e.g. *i am no_RB longer affiliated* **Determiner**: e.g. *there's no_DT money* **Response Marker**: e.g. *no_RE but i can make a chick break for you (.)*
okay	**Adjective**: e.g. *is this okay_JJ for everyone* **Adverb**: e.g. *we're doing okay_RB* **Response marker**: e.g. *okay_RE, I'll do it.*

Individual token(s)	Tagging practice
so	**Meaning "so that" and "therefore"**: Tag **IN**, e.g. *[first name1] will be there so_IN he will have the occasion to speak out there* **Adverbial use: Tag RB**, e.g. *so_RB good* **In certain fixed expressions: Tag RB**, e.g. *or so_RB, and so_RB on* **Clause-final** or **not related to main clause: Tag DM,** e.g. *pedagogical way so_DM _0*
so that	**Subordinating conjunction**: e.g. *so_IN that_IN WE do not go to the politicians* **Discourse Marker, followed by Determiner**: e.g. *so_DM that_DT's strong*
such	**Predeterminer:** e.g. *such_PDT a darling* **Determiner:** e.g. *such_DT documents*
that	**Determiner**: e.g. *put that_DT thing away; that_DT came after the mcsherry re-reform* **Relative pronoun**: e.g. *first thing that_PRE crosses your mind* **Subordinating conjunction**: e.g. *the problem was that_IN she was like RUNNING*
the -er the -er	*the_DT broader_JJR the_DT better_JJR*
the same	*the_DT same_NN* (if no noun is following)
though	**Conjuction**: e.g. *though_IN we are prepared* (*even_IN though_IN*, cf. 6.2. VOICE List of Multi-words) **Adverb**: e.g. *you know what's funny though_RB*
to	**Infinitive use**: e.g. *to_TO go* **Preposition:** e.g. *to_IN the market*
up to	*it's up_RB to_IN the labor market, up_RB to_IN fifteen minutes* (vs. *up_JJ to_JJ date_JJ*, cf. 6.2. VOICE List of Multi-words)
use(d) (to)	**Adjective**: e.g. *i'm used_JJ to it; a used_JJ car* **Verb**: e.g. *I used_VVD to do something; the designers have used_VVN that*

4 The VOICE lemmatization guidelines

This section provides general information on how the VOICE tagging formats were treated in the lemmatization process, followed by an explanation of the lemmatization rules for individual categories, listed in alphabetical order.

4.1 Lemmatization and VOICE Tagging formats

In VOICE, all tokens are assigned a maximum of two tag combinations in the basic format FORM(FUNCTION). However, due to this format, and the format for tag ambiguities, each basic tag can consist of more than one individual tag (cf. 3.2. Tagging formats in VOICE). Hence, each basic tag can be assigned more than one lemma. The rules for lemmatization are as follows: If the **lemmata** of the individual tags **converge**, only one lemma is assigned, e.g. token: *rules_NNS/VVZ*, lemma: *rule*. In those cases where the **lemmata** for the individual tags **do not converge**, more than one lemma is assigned. This was often the case for ambiguities, e.g. token: *including_IN/VVG*, lemmata*: including, include,* and with tokens that were assigned different tags for form and function, e.g. token: *feeling_VVG(NN)*, lemmata: *feel, feeling;* token: *preserved_V(JJ)*, lemmata: *preserve, preserved.*

4.2 Contracted forms

Contracted forms are assigned the corresponding lemma of their full forms, e.g. token: *'ve* (e.g. in *you've*), lemma: *have*; token: *'re* (e.g. in *you're*), lemma: *be*, token: *n't* (e.g. in *don't*), lemma: *not* etc.[6] The tokens *ta* (in *gotta*) and *na* (in *gonna*, *wanna)* are assigned the lemma *to*.

4.3 Interjections

The lemma for **interjections** (tagged with UH) is identical with the token itself, e.g. token: *yeah*, lemma: *yeah* (not e.g. *yes*).

4.4 Nouns

The lemmata for **nouns** are their respective singular forms, e.g. token: *languages*, lemma: *language*, the lemma for adjectives their positive form, e.g. token: *bigger,* lemma*: big*.

Pluralia tantum are not reduced to a non-existent singular form, e.g. token: *trousers,* lemma: *trousers*.

For all items in the closed list of **compound nouns for VOICE** (cf. 6.4), each part of the compound receives a separate lemma, e.g. *business cards*: token: *business, cards*, lemmata: *business, card.* All parts of these compound nouns are considered to function as a noun unit. Forms which are part of a noun compound but do not have a noun form, are not reduced to their base forms but lemmatized in their inflected form, although in other co-texts they might belong to another lemma, e.g. token: *added value*, lemmata: *added, value (not: add, value)*

4.5 Numbers

Ordinal numbers, e.g. *seventh* are lemmatized as such and not reduced to their cardinal form, e.g. token: *seventh,* lemma: *seventh*.

[6] Exception for genitive –'s: lemma = 's

4.6 Pronouns

Pronouns, i.e. objective (e.g. *me, you, him*, ...), reflexive (e.g. *myself, yourself, himself*, ...) and possessive (e.g. mine, yours, his), as well as possessive determiners (*my, your, his*, …) are assigned their nominative forms as lemma, e.g. token: *your*, lemma: *you*.

4.7 Quantifiers

For the **quantifiers** *much, many, more, most, less, lesser* and *least*, the lemmata are identical with their respective forms (no reduction to a positive form), e.g. token*: least,* lemma*: least* (not: *little* or *less*)

4.8 Verbs

For **verbs** the lemma is identical with the base form. For example, the tokens *go, goes, going, went, gone* and *gon* (in *gonna*) constitute a single inflectional paradigm and are all assigned the lemma *go*.

Modal verbs are lemmatized according to the same principle as verbs, i.e. the lemma is the respective base form. Examples: token: *can*, *ca* (in *can't*), lemma: *can*; token: *shall*, *should*, lemma: *shall*; token: *will, would, wo('nt),* lemma: *will*.

5 Sources

Baayen, R. Harald; Piepenbrock, Richard; Gulikers, Leon. 1995. "The CELEX Lexical Database (CD-ROM)". Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Biber, Douglas (ed.). 1999. *Longman grammar of spoken and written English*. (1st edition). Harlow: Longman.

Brants, Thorsten. 2000. "TnT - A Statistical Part-of-Speech Tagger". In. *Proceedings of the Sixth Applied Natural Language Processing Conference*. Seattle, WA, 224-231.

Breiteneder, Angelika; Pitzl, Marie-Luise; Majewski,Stefan; Theresa Klimpfinger. 2006. "VOICE recording - Methodological challenges in the compilation of a corpus of spoken ELF". *Nordic Journal of English Studies* 5(2), 161–188. http://hdl.handle.net/2077/3153 (16 April 2014)

Brill, Eric. 1995. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging". *Computational Linguistics* 21(4), 543-565.

Brill, Eric; Wu, Jun. 1998. "Classifier Combination for Improved Lexical Disambiguation". In. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Volume 1. Association for Computational Linguistics. Stroudsburg, PA (1), 191-195.

Carter, Ronald; McCarthy, Michael. 2006. *Cambridge Grammar of English. CD-Rom. A comprehensive guide to spoken and written English usage.* Cambridge: CUP.

Daelemans, Walter; Zavrel, Jakub; Berck, Peter; Gillis, Steven. 1996. "MBT: A memory-based part of speech tagger generator". In. *Proceedings of the Fourth Workshop on Very Large Corpora*. Copenhagen, 14-27.

Daume III, Hal; Kumar, Abhishek; Saha, Avishek. 2010. "Frustratingly easy semi-supervised domain adaptation". In. *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing.* ACL 2010. Uppsala, 15 July 2010, 53-59.

Davies, Mark. n.d. "Word frequency lists and dictionary from the Corpus of Contemporary American English". http://www.wordfrequency.info/free.asp (29 November 2012).

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Giménez, Jesús; Marquez, Lluis. 2003. "Fast and accurate part-of-speech tagging: The SVM approach revisited". In Nicolov, Nicolas (ed.). *Recent Advantages in Natural Language Processing III: Selected papers from RANLP 2003.* Samokov, Bulgaria. Amsterdam: Benjamins, 153-163.

Hornby, Albert Sydney; Wehmeier, Sally; McIntosh, Colin; Turnbull, Joanna; Ashby, Michael. 2007. *Oxford advanced learner's dictionary*. (7th edition). Oxford: Oxford University Press.

Kilgarriff, Adam. 2006. "BNC database and word frequency lists". http://www.kilgarriff.co.uk/bnc-readme.html (29 November 2012).

Lafferty, John; McCallum, Andrew; Pereira, Fernando. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In. *Proceedings of International Conference in Machine Learning (ICML-01)*. Williamstown, MA, 282-289.

Lager, Torbjörn. 2001. "Transformation-Based Learning of Rules for Constraint Grammar Tagging". In. *Proceedings of the 13th Nordic Conference in Computational Linguistics*. Uppsala.

Linguistic Data Consortium (LDC). 1999. "Addendum to the part-of-speech tagging guidelines for the Penn Treebank project (Modifications for the SwitchBoard corpus)". http://www.cis.upenn.edu/~bies/manuals/tagguid2.pdf (18 October 2012).

Marcus, M.P; Marcinkiewicz, M.A; Santorini, B. 1993. "Building a large annotated corpus of English: The Penn Treebank". *Computational Linguistics* 19(2), 313-330.

Miller, George A. 1995. "WordNet: A Lexical Database for English". *Communications of the ACM* 38(11), 39-41.

Nelson, Gerald. 2005. The ICE Tagging Manual. Revised version. http://ice-corpora.net/ice/taggingmanual.doc (16 April 2014)

Osimk-Teasdale, Ruth.2013. "Applying existing tagging practices to VOICE". In Mukherjee, Joybrato; Huber, Magnus (eds.). *Corpus linguistics and variation in English: Focus on Nonnative Englishes (Proceedings of ICAME 31)*. Helsinki: VARIENG.

Osimk-Teasdale, Ruth. in prep. *Parts of speech in English as a lingua franca: the POS tagging of VOICE.* PhD Thesis, University of Vienna.

Pitzl, Marie-Luise; Breiteneder, Angelika; Klimpfinger, Theresa. 2008. "A world of words: processes of lexical innovation in VOICE". *Views* 17, 21-46.

Quirk, Randolph (ed.). 1997. *A comprehensive grammar of the English language*. (14th edition). London [u.a.]: Longman.

Radeka, Michael. 2009. *Paralleles transformationsbasiertes Lernen: Kombination von Regelmengen mit einem korpusbasierten Selektionsverfahren*. Magisterarbeit, Ruprecht-Karls-Universität Heidelberg.

Radeka, Michael. in prep. *Tagging VOICE: A parallel stacked TBL approach.* PhD Thesis, University of Vienna.

Ratnaparkhi, Adwait. 1996. "A maximum entropy part-of-speech tagger". In. *Proceedings of the First Conference on Empirical Methods in NLP*. Philadelphia, PA, 133-142.

Santorini, Beatrice. 1991. "Part of Speech Tagging Guidelines for the Penn Treebank Project". http://www.personal.psu.edu/xxl13/teaching/sp07/apling597e/resources/Tagset.pdf (18 October 2012).

Santorini, Beatrice. 1995. "Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd Printing)". http://www.ling.helsinki.fi/kit/2010s/clt236/docs/PennTaggingGuide.pdf (3 December 2012).

Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In. *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, UK, 44-49.

Seidlhofer, Barbara. 2011. *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Shen, Libin; Satta, Giorgio; Joshi, Aravind K. 2007. "Guided learning for bidirectional sequence classification". In. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*. Prague, 760-767.

Toutanova, Kristina; Klein, Dan; Manning, Christopher; Singer, Yoram. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network". In. *Proceedings of HLT-NAACL 2003*. Edmonton, Canada, 252-259.

van Halteren, Hans; Daelemans, Walter; Zavrel, Jakub. 2001. "Improving accuracy in word class tagging through the combination of machine learning systems". *Computational Linguistics* 27/2, 199-229.

VOICE Project. 2007a. "Spelling conventions". http://www.univie.ac.at/voice/documents/VOICE_spelling_conventions_v2-1.pdf (19 May 2011).

VOICE Project. 2007b. "VOICE Transcription Conventions [2.1]". http://www.univie.ac.at/voice/voice.php?page=transcription_general_information (6 December 2012).

VOICE Project. 2007c. "Mark-up conventions. VOICE Transcription Conventions [2.1]". http://www.univie.ac.at/voice/documents/VOICE_mark-up_conventions_v2-1.pdf (18 July 2011).

Volk, Martin; Schneider, Gerold. 1998. "Adding Manual Constraints and Lexical Look-up to a Brill-Tagger for German". In. *Proceedings of the ESSLLI-Workshop on Recent Advances in Corpus Annotation*. Saarbrücken.

Wu, Dekan; Ngai, Grace; Carpuat, Marine. 2004. "Raising the Bar: Stacked Conservative Error Correction Beyond Boosting". In. *Proceedings of the fourth International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon, 21-24.

6 Appendix

6.1 VOICE List of Formulaic Items

DESCRIPTION: Contains formulaic expressions which are in the closed list of **VOICE Formulaic Items**, including greetings, farewells, please, thanks, apologies, wishes and miscellaneous expressions. It is based on the exclamations listed in OALD7 and the categorization of formulaic expressions in ICE (Nelson 2005: 13).

TAGGING: **Tag:** FI

The list of formulaic items generally only contains very short formulaic chunks. In slightly longer, syntactically analysable stretches, the syntactic co-text was not tagged with FI but with the 'conventional' POS tag. e.g. *thank_FI you_FI very_RB much_RB*.

greetings & farewells
bye
bye-bye
ciao
good afternoon
good day
good evening
good morning
goodbye
goodnight
hello
hey
hi
see you
welcome

please & thanks
thanks
thank you
please
you're welcome

apologies
pardon
pardon me
sorry
excuse me

expletives
christ
damn
dammit
dear
dude
fuck
gee
gosh
heck
jesus
(my) god
(my) goodness
shit
shoot
boy
man

wishes
congratulation(s)
happy birthday
happy ramadan
merry christmas

miscellaneous expressions
attention
bingo
bravo
cheers (not used to mean '*thanks*' in VOICE)
encore
viva

6.2 VOICE List of Multi-words

DESCRIPTION: Included in this **closed list** are the most frequent multi-word chunks from the word classes **Adverb, Adjective, Conjunction** and **Preposition** in VOICE. This list is based on Multi-word items in the OALD7 (reference dictionary), those used in the BNC and those which appear on the VOICE bi- and trigrams. Also included are foreign Multi-word items, mostly of Latin or Greek origin (e.g. *ad hoc*).

TAGGING: Each part of the multi-word item is assigned the **same tag**, e.g. *a_RB lot_RB*
NB: Many of the items listed here have counterparts which do not function as a multi-word and receive 'conventional' POS tags. These cases have been disambiguated, e.g. *so_IN that_IN when we go back* vs. *so_DM that_DT 's why*; *kind_RB of_RB good* vs. *that kind_NN of_DT law*.

a) ADVERBS (Tags: RB)

a bit
a cappella
a little
a little bit
a lot
a priori
ad hoc
ad nauseam
all right
any more
as well
at all
at least
de facto
et cetera
for example
for instance
for sure
in general
kind of
more or less
of course
over there
per capita
per cent
per se
sort of
sui generis
vice versa

b) ADJECTIVES (Tags: JJ)

a cappella
a priori
ad hoc
ad nauseam
all right
brand new
de facto
far eastern
fed up
fully fledged
in vitro
middle eastern
new age
next door
number one
out of date
per capita
per cent
per cent
politically correct
roman catholic
social democratic
sold out
sui generis
up to date
upper class
well balanced
well built
well defined
well developed
well disposed
well done
well informed
well known
well paid
well used
worked up

c) CONJUNCTIONS (Tags: IN)

even if
even though
now that
so that

d) PREPOSITIONS (Tags: IN)

according to
because of
depending on
in order to
in terms of
instead of
next to
out of
such as
vis-à-vis

6.3 VOICE List of Discourse Markers and Interjections

6.3.1 Single word discourse markers

Items: *like, look, whatever, well, so, right*
Tag: DM

6.3.2 Multi-word discourse markers

Items: *I mean, I see, mind you, you know, you see*
Tags: Multi-word discourse markers are tagged with a conventional tag for form, and the tag DM for function, for all parts of the discourse marker:

I_PP(DM) mean_VVP(DM)
I_PP(DM) see_VVP(DM)
mind_VVP(DM) you_PP(DM)
you_PP(DM) know_VVP(DM)
you_PP(DM) see_VVP(DM)

6.3.3 Interjections

DESCRIPTION: These are the items listed as discourse markers in the VOICE Mark-up conventions (VOICE Project 2007c: 4). They do not have a homonym in a different word class category but fulfil the following discourse functions. The items in green have been added for VOICE POS.

TAGGING: **Tag:** UH (NB: Non-English discourse markers are tagged **FW**.)

Interjection	Function
er, erm	**Hesitation**/filler
huh	tag-question
yay, yipee, whoohoo, mm:	**Exclamations** joy/enthusiasm
haeh	questioning/doubt/disbelief
a:h, o:h, wow, poah	astonishment/surprise
oops	apology
ooph	exhaustion
ts, pf	disregard/dismissal/contempt
ouch, ow	pain
sh, psh	requesting silence
oh-oh:, u:h	anticipating trouble
ur, yuck	disapproval/disgust
oow	pity/disappointment
blah	**expressing predictability or lack of interest for something**

6.4 VOICE List of Compound Nouns

DESCRIPTION: The VOICE list of Compound Nouns is a closed list, consisting of 1) items listed as multiword noun units in our **reference dictionary**, e.g. *public service*, but also e.g. *master of ceremony* and 2) multiword noun units occurring **most frequently** (30 times or more) in the **n-gram lists** for our data, e.g. *joint program*. This list includes all 359 noun combinations tagged as compound nouns in VOICE, however, not including those where the first word ends in the morpheme *–ing* (e.g. *swimming pool*) (see 3.4.1.3.)

TAGGING: Format for **singular** compound nouns: e.g. *academic_NN year_NN*
Format for **plural** compound nouns: e.g. *business_NNS cards_NNS*

academic year
added value
adult education
affirmative action
age group
age limit
alarm clock
amusement park
armed forces
art gallery
art history
art nouveau
artificial intelligence
artificial language
assistant professor
associate professor
au pair
auxiliary language
back door
balance sheet
bank holiday
banoffi pie
bar code
best practice
big bang
big toe
birth rate
black box
black market
black sheep
blister pack
blood pressure
body language
bonded warehouse
bonfire night
bottom line
brain drain
brand name
bullet point
bus stop
business administration
business card
business school
calendar year
capital city
carbon copy
cash and carry
cash flow
catchment area
central bank
central government
chain reaction
checks and balances
chief executive
chip card
christmas tree
civil rights
civil servant
civil war
clean up
coat of arms
code of practice
coffee break
coffee shop
common denominator
common ground
common law
common market
common room
common sense
community service
computer science
conceptual art
condensed milk
consumer goods
contact person
continental shelf
convenience store
court of appeal
credit card
critical mass
culture shock
current account
dance floor
day off
day out
dead end
department store
differential equation
digestive system
direct action
double agent
double room
dress code
dress rehearsal
due date

duty-free shop
end product
end result
education system
exchange rate
exclamation mark
extreme sports
fairy tale
family name
feta cheese
field trip
financial aid
fire alarm
fire station
first floor
first language
first name
fish and chips
flat rate
flip chart
flow chart
focal point
focus group
frame of reference
free trade
free will
front desk
front page
front runner
full professor
further education
general assembly
general knowledge
general practice
general public
generation x
giant slalom
global village
golden retriever
golden rule
good faith
good practice
good sense
grand master
grass roots
green light
grocery store
gross domestic product
ground floor
ground rule
group work
half day
hard copy
head office
health care
heart attack
high five
high heels
high school
high season
high water
higher education
home economics
home page
human nature
human resources
ice cream
income support
information technology
intelligence test
internal market
inverted commas
irish coffee
ivory tower
jet lag
job description
joint degree
joint master
joint program
joint venture
labor force
labor market
last name
lead time
legal action
legal tender
letter of intent
life expectancy
life insurance
lingua franca
local government
loose end
low point
low tide
lower house
lump sum
main street
market share
master of ceremonies
master plan
maternity leave
media studies
member state
membership criterion
middle ages
middle class
middle name
middle school
military service
mineral water
minimum wage
minority government
mission statement
mobile phone
money-back guarantee
monopoly money
mother tongue
movie theater
mutual recognition
nation state
national anthem
national curriculum
native speaker
natural gas
natural science
news agency
normal distribution
nuclear physics
nuclear power
office hours
old age
open market

open season
order of magnitude
organic chemistry
paper cutter
peace process
pencil case
pension scheme
petrol station
phone call
phone number
point of view
point of sale
political correctness
political science
political scientist
population explosion
position paper
post office
present day
press agency
press conference
press release
press secretary
pressure cooker
price controls
price tag
price war
primary school
prime minister
private company
private law
private school
private sector
production line
public access
public opinion
public relations
public school
public sector
public service
public transport
public transportation
purchase price
quality assurance
quality control
question mark
race car
raw material
red carpet
red wine
red-light district
research and development
response time
right wing
road map
rock music
role model
round trip
sales rep
science fiction
score sheet
seat belt
second language
second name
secondary school
security council
seed money
senior citizen
service provider
short cut
short time
side street
sim card
sine qua non
ski lift
small talk
social fund
social inclusion
social science
social security
social studies
social worker
split second
stainless steel
star sign
state university
status quo
stock exchange
stock market
student union
stream of
consciousnesssuccess story
summer school
supply and demand
suspension bridge
swiss cheese
tape recorder
target language
task force
telephone number
terms of reference
three quarters
time bomb
time frame
time limit
time span
time zone
top ten
town hall
track record
trade union
trash can
travel agency
trust fund
tuition fees
upper class
vested interest
video camera
voluntary service
voluntary work
way out
web page
welfare state
white fish
white wine
wine gum
work experience
work permit
world cup
youth hostel
youth organization