Annotating tense, mood and voice for English, French and German

Anita Ramm, Sharid Loáiciga, Annemarie Friedrich, Alexander Fraser

Angaben zur Veröffentlichung / Publication details:

Ramm, Anita, Sharid Loáiciga, Annemarie Friedrich, and Alexander Fraser. 2017. "Annotating tense, mood and voice for English, French and German." In *Proceedings of ACL 2017, System Demonstrations, July 30 - August 4, 2017, Vancouver, Canada*, edited by Mohit Bansal and Heng Ji, 1–6. Stroudsburg, PA: Association for Computational Linguistics. https://aclanthology.org/P17-4001.

Annotating tense, mood and voice for English, French and German

Anita Ramm[1,4] **Sharid Loáiciga**[2,3] **Annemarie Friedrich**[4] **Alexander Fraser**[4]
[1]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
[2]Département de Linguistique, Université de Genève
[3]Department of Linguistics and Philology, Uppsala University
[4]Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximilians-Universität München
ramm@ims.uni-stuttgart.de sharid.loaiciga@unige.ch
{anne,fraser}@cis.uni-muenchen.de

Abstract

We present the first open-source tool for annotating morphosyntactic tense, mood and voice for English, French and German verbal complexes. The annotation is based on a set of language-specific rules, which are applied on dependency trees and leverage information about lemmas, morphological properties and POS-tags of the verbs. Our tool has an average accuracy of about 76%. The tense, mood and voice features are useful both as features in computational modeling and for corpus-linguistic research.

1 Introduction

Natural language employs, among other devices such as temporal adverbials, *tense* and *aspect* to locate situations in time and to describe their temporal structure (Deo, 2012). The tool presented here addresses the automatic annotation of *morphosyntactic tense*, i.e., the tense-aspect combinations, expressed in the morphology and syntax of verbal complexes (VC). VCs are sequences of verbal tokens within a verbal phrase. We address German, French and English, in which the morphology and syntax also includes information on mood and voice. Morphosyntactic tenses do not always correspond to *semantic tense* (Deo, 2012). For example, the morphosyntactic tense of the English sentence "He is leaving at noon." is *present progressive*, while the semantic tense is *future*. In the remainder of this paper, we use the term *tense* to refer to the morphological tense and aspect information encoded in finite verbal complexes.

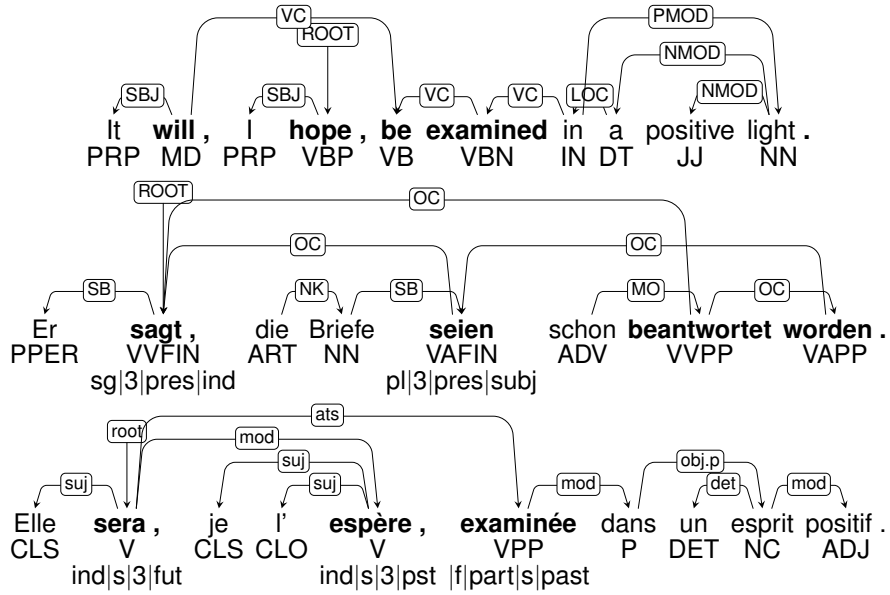
Corpus-linguistic research, as well as automatic modeling of mono- and cross-lingual use of tense, mood and voice will strongly profit from a reliable automatic method for identifying these clausal features. They may, for instance, be used to classify texts with respect to the epoch or region in which they have been produced, or for assigning texts to a specific author. Moreover, in cross-lingual research, tense, mood, and voice have been used to model the translation of tense between different language pairs (Santos, 2004; Loáiciga et al., 2014; Ramm and Fraser, 2016)). Identifying the morphosyntactic tense is also a necessary prerequisite for identifying the semantic tense in synthetic languages such as English, French or German (Reichart and Rappoport, 2010). The extracted tense-mood-voice (TMV) features may also be useful for training models in computational linguistics, e.g., for modeling of temporal relations (Costa and Branco, 2012; UzZaman et al., 2013).

As illustrated by the examples in Figure 1, relevant information for determining TMV is given by syntactic dependencies and partially by part-of-speech (POS) tags output by analyzers such as Mate (Bohnet and Nivre, 2012). However, the parser's output is not sufficient for determining TMV features; morphological features and lexical information needs to be taken into account as well. Learning TMV features from an annotated corpus would be an alternative; however, to the best of our knowledge, no such large-scale corpora exist.

A sentence may contain more than one VC, and the tokens belonging to a VC are not always contiguous in the sentence (see VCs A and B in the English sentence in Figure 1). In a first step, our tool identifies the tokens that belong to a VC by analysing their POS tags as well as the syntactic dependency parse of the sentence. Next, TMV values are assigned according to language specific hand-crafted sets of rules, which have been developed based on extensive data analysis. The system contains approximately 32 rules for English and 26 rules for German and for French. The TMV values are output along with some additional in-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 1–6
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-4001

(1) Output of MATE parser:



(2) Extraction of verbal complexes based on dependencies;
(3) Assignment of TMV features based on POS sequences, morphological features and lexical rules:

A	will be examined	MD[will] VB[be] VBN	→ **futureI**	**indicative**	**passive**
B	hope	VBP	→ **present**	**indicative**	**active**
C	sagt	VFIN[pres/ind]	→ **present**	**indicative**	**active**
D	seien beantwortet worden	VAFIN[pres/ind] VVPP VVPP[worden]	→ **present**	**indicative**	**passive**
E	sera examinée	V[ind/fut] VPP[part/past]	→ **futureI**	**indicative**	**passive**
F	espère	V[ind/pst]	→ **present**	**indicative**	**active**

Figure 1: **Example** for TMV extraction.

formation about the VCs into a TSV file which can easily be used for further processing.

Related work. Loáiciga et al. (2014) use rules to automatically annotate tense and voice information in English and French parallel texts. Ramm and Fraser (2016) use similar tense annotation rules for German. Friedrich and Pinkal (2015) provide a tool which, among other syntactic-semantic features, derives the tense of English verbal complexes. This tense annotation is based on the set of rules used by Loáiciga et al. (2014)

For English, PropBank (Palmer et al., 2005) contains annotations for tense, aspect and voice, but there are no annotations for subjunctive constructions including modals. The German TüBa-D/Z corpus only contains morphological features.[1]

Contributions. To the best of our knowledge, our system represents the first open-source[2] system which implements a reliable set of derivation rules for annotating tense, mood and voice for English, French and German. Furthermore, the online demo[3] version of the tool allows for fast text processing without installing the tool.

2 Properties of the verbal complexes

In this section, we describe the morphosyntactic features that we extract for verbal complexes.

2.1 Finite and non-finite VCs

We define a verbal complex (VC) as a sequence of verbs within a verbal phrase, i.e. a sentence may include more than one VC. In addition to the verbs, a VC can also contain verbal particles and negation words but not arguments. We distinguish between finite VCs which need to have at least one finite verb (e.g. "sagt" in Figure 1), and non-finite VCs which do not; the latter consist of verb forms such as gerunds, participles or infinitives (e.g. "to support"). Infinitives in English and German have to occur with the particles *to* or *zu*, respectively,

[1] http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html

[2] https://github.com/aniramm/tmv-annotator

[3] https://clarin09.ims.uni-stuttgart.de/tmv/

while in French, infinitives may occur alone. We do not assign the TMV features to non-finite VCs. Our tool marks finiteness of a VC using a binary feature "yes" (finite) and "no" (non-finite).

2.2 Tense, mood, voice

The identification of TMV features for a VC requires the analysis of lexical and grammatical information, such as inflections, given by the combination of verbs. For example, the English *present continuous* requires the auxiliary *be* in present tense and the gerundive form of the main verb (e.g. "(I) am speaking").

Mood refers to the distinction between *indicative* and *subjunctive*. Both of these values are expressed in the inflection of finite verbs in all the considered languages. For example, the English verb "shall" is indicative, while its subjunctive form is "should." In English, tense forms used in subjunctive mood are often called *conditionals*; for German, they are referred to as *Konjunktiv*.

Voice differentiates between *active* and *passive* constructions. In all three languages, the passive voice can be recognized by searching for a specific verb. For example, the passive voice in English requires the auxiliary ***be*** in a tense-specific form, e.g., "(I) am **being** seen" for *present progressive* or "(he) has **been** seen" for *present perfect*.

Details on how our tool automatically identifies TMV features will be described in Section 3.

2.3 Negation

VCs may include negation. Our tool outputs a binary negation value to VCs depending on whether a negation word (identified by checking for a language-specific POS-tag) is part of the verbal dependency chain. If a negation exists, the feature value is "yes", and "no" otherwise.

2.4 Main verb

Within a VC, the main verb bears the semantic meaning. For example, in the English VC "would have read," the participle "read" is considered to be the main verb. The main verb feature may contain a single verb or a combination of a verb with the verb particle. In the following, we describe the detection of the main verbs for each of the three languages under consideration.

English and French. In English and French VCs, the very last verb in the VC is considered to be the main verb. For example, in the English VC "will be examined", "examined" is marked as the main verb. Verb particles are considered as a part of the main verb and are attached to the corresponding main verb, e.g., the main verb of the non-finite English VC "to move up" is "move-up."

German In general, the main verbs in German have specific POS-tags (*VV**) (see, for example, (Scheible et al., 2013)). In most German VCs, there is only one verb with such a POS-tag. However, there are a few exceptions. For example, the *recipient passive* is built with full verbs *bekommen, kriegen*, as well as *lernen, lassen, bleiben* and an additional meaning-bearing full verb. Thus, in such constructions, there are two verbs tagged as VV* (e.g. "Ich bekomme$_{VVFIN}$ das Buch geschenkt$_{VVPP}$." ("I receive the book donated")). Recipient verbs are not treated as main verbs if they occur with an additional full verb. In case there are no verbs tagged with *VV**, the last verb in the chain is considered to be the main verb.

3 Deriving tense, mood and voice

In this section, we give a short overview of the methods used to derive TMV information.

3.1 Extraction of VCs

The tokens of a VC are not necessarily contiguous. They may be separated by a coordination, adverbials, etc., or even include nested VCs as in Figure 1. This makes it necessary to take syntactic dependencies into account. The extraction of VCs in our tool is based on dependency parse trees in the CoNLL format.[4] The first step is the identification of all VC-beginning elements v_b within a sentence, which include finite verbs (English, French and German) and infinitival particles (English, German). They are identified by searching for specific POS-tags. For each v_b, the remaining elements of the VC are collected by following the dependency relations between verbs. Consider for example the finite verb "will" in Figure 1. It is identified as a v_b because of its POS tag *MD*. We now follow the dependency path from "will" to "be" and from "be" to "examined". The resulting VC is thus "will be examined."

[4] In this work, we use the Mate parser for all three languages. https://code.google.com/archive/p/mate-tools/wikis/ParserAndModels.wiki.

finite	mood	tense	voice	example (active voice)
yes	ind	present	act	(I) work
		presProg		(I) am working
		presPerf		(I) have worked
		presPerfProg		(I) have been working
		past		(I) worked
		pastProg		(I) was working
		pastPerf		(I) had worked
		pastPerfProg		(I) have been working
		futureI	pass	(I) will work
		futureIProg		(I) will be working
		futureII		(I) will have worked
		futureIIProg		(I) will have been working
	subj	condI		(I) would work
		condIProg		(I) would be working
		condII		(I) would have worked
		condIIProg		(I) would have been working
no	-	-	-	to work

Table 1: TMV combinations for English.

finite	mood	tense	voice	example (active voice)
yes	ind	present	act pass	(je) travaille
		presPerf		(je) viens de travailler
		perfect		(j'')ai travaillé
		imperfect		(je) travaillais
		pastSimp		(je) travaillai
		pastPerf		(j'')eus travaillé
		pluperfect		(j'')avais travaillé
		futureI		(je) travaillerai
		futureII		(j'')aurai travaillé
		futureProc		(je) vais travailler
	subj	present		(je) travaille
		past		(j'')aie travaillé
		imperfect		(je) travaillasse
no	-	-	-	travailler

Table 2: TMV Combinations for French.

finite	mood	tense	voice	example (active voice)
yes	ind	present	act pass	(ich) arbeite
		perfect		(ich) habe gearbeitet
		imperfect		(ich) arbeitete
		pluperfect		(ich) hatte gearbeitet
		futureI		(ich) werde arbeiten
		futureII		(ich) werde gearbeitet haben
	konjI konjII	present		(er) arbeite/arbeitete
		past		(er) habe/hätte gearbeitet
		futureI+II		(er) würde arbeiten / gearbeitet haben
no	-	-	-	zu arbeiten

Table 3: TMV combinations for German.

3.2 TMV extraction rules

English. The rules for English make use of the combinations of the *functions* of the verbs within a given VC. Such functions are for instance *finite verb* or *passive auxiliary*. According to the POS combination of a VC and lexical information, first, the function of each verb within the VC is determined. Subsequently, the combination of the derived functions is mapped to TMV values. For example, the following functions will be assigned to the verbs of the VC "will be examined" in Figure 1: "will" → *finite-modal*, "be"→ *passive-auxiliary*, "examined" → *past-participle*. This particular combination of verb functions leads to the TMV combination *futureI/indicative/passive*. Table 1 contains the set of possible TMV combinations that our tool extracts for English.

French. The rules for French are defined on the basis of the reduction of the verbs to their morphological features. The morphological features of the verbs are derived from the morphological analysis of the verbs, as well as their POS-tags. The rules specify TMV values for each of the possible sequences of the morphological features. For example, the VC "sera examinée" is mapped to the morphological feature combination *V-indfut-V-partpast* which, according to our rule set, leads to the TMV *futureI/indicative/passive*. In some cases, the lexical information is used to decide between ambiguous configurations. For example, some *perfect/active* forms are ambiguous with *present/passive* forms. For instance, "Jean est parti" and "Jean est menacé" are both composed of the verb "est" + past participle, but they have different meaning: "Jean has left" vs. "Jean is threatened." Information about the finite verb helps to distinguish between the two constructions. Table 2 shows the French TMV combinations.

German. The rules are based on POS tags, morphological analysis of the finite verbs and the lemmas of the verbs. We group the rules by the number of tokens contained in the VC, as we have observed that each combination of TMV features requires a particular number of tokens in the VC. For each length, we specify which tense and mood of the finite verb lead to a specific TMV. Similarly to French, in some contexts, we need to use lexical information to decide on TMV.

Take for example the VC "seien beantwortet worden" from Figure 1. Its POS sequence is *VAFIN-VVPP-VAPP*, so we use rules defined for the POS length of 3. We first check the mood of the finite verb "seien" which is *subj* (subjunctive). The combination of *subj* with the morphological tense of the finite verb *pres* leads to the mood value *konjunktivI* and the tense value *past*. As the verb *werden*, which is used for passive constructions in German, occurs in the VC, we derive the voice value *passive*. Thus, the resulting annotation is *past/konjunktivI/passive*. Table 3 shows TMV value combinations for German.

3.3 Extraction of voice

In all three languages, it is difficult to distinguish between stative passive and tenses in the active voice. For instance, the German VCs "ist geschrieben (is written)" and "ist gegangen (has gone)" are both built with the auxiliary *sein* and a past participle. The combination of POS tags is same for both cases, and the morphological features of the finite verb (*pres/ind*) correspond to the German perfect tense in active voice. This, however, holds only for verbs of movement and a few other verbs. Verbs such as "schreiben (to write)" are in this specific context *present/passive* (stative passive in present tense) and not *perfect/active* which is the case for the VC "ist gegangen".

To disambiguate between these constructions, we use a semi-automatically crafted list of the German and French verbs that form *perfect/active* with the auxiliary *sein/être (be)* instead of *haben/avoir (have)*, which is used for the majority of the verbs. We extract these lists from different corpora by counting how often verbs occur with *sein/haben* and *être/avoir*, respectively. We manually validate the resulting verb lists.

When a VC with a POS sequence that is ambiguous in the above explained way is detected, we check whether the main verb is in the list of "sein/être" verbs. If that is the case, the corresponding active tense is annotated. Otherwise, the VC is assigned the corresponding passive tense.

In the case of English, the disambiguation is somewhat easier. To differentiate between "is written" and "has written," we use information about the finite verb within the VC. In the case where we have *be*, we assume to have passive voice in combination with an appropriate tense. In case of *have*, the voice is active.

4 Annotation tool

The tool is implemented in Python. It takes as input the parsed text file in the CoNLL format. For the rule development, as well as evaluation, we used the Mate parser (Bohnet and Nivre, 2012), which can be applied on all of the three languages addressed here. For German and French, we use the joint model for parsing, tagging and morphological analysis including lemmatization. For English, only tagging and parsing is required. In general, the TMV annotation tool is applicable on the output of arbitrary parsers as long as their models use the same POS- and dependency tags as Mate.

The tool outputs a TSV file with TMV annotations. An example output is shown in Table 4. The columns are specified as follows: sentence number, indices of the elements of a VC separated by a comma, elements of a VC separated by a comma, finite, main verb (if more than one, separated by a comma), tense, mood, voice, progressive (only for English), coordination and negation. The German TSV output has an additional column with boundaries of a clause in which a VC is placed.[5] We additionally provide a script for the conversion of the annotations into HTML format which allows for quick and easy examination of the annotations.

5 Evaluation

We manually evaluate annotations for 157 German VCs, 151 English Vcs and 137 French VCs extracted from a set of randomly chosen sentences from Europarl (Koehn, 2005). The results are shown in Table 5.

Language	tense	mood	voice	all
EN	81.5	88.1	86.1	76.8
DE	80.8	84.0	81.5	76.4
FR	86.1	93.4	82.5	75.2

Table 5: Accuracy of TMV features according to manual evaluation.

For French, the overall acurracy is 75%, while the accuracy of German and English annotations is 76%. Based on the manually annotated sample, we estimate that 23/59/85% (for EN/DE/FR) of the erroneous annotations are due to parsing errors. For instance, in the case of English, the VC extraction process sometimes adds gerunds to the VC and interprets them as a present participle. Similarly, for French, a past participle is added, which erroneously causes the voice assignment to be passive. Contrary to German and English, French has higher mood accuracy, since mood is largely encoded unambiguously in the verb morphology. For German, false or missing morphological annotation of the finite verbs causes some errors, and there are cases not covered by our rules for identifying stative passive.

Our rule sets have been developed based on extensive data analysis. This evaluation presents a

[5]The clause boundary identification is based on the sentence punctuation (e.g. comma, colon, semicolon, hyphen, etc). For more sophisticated clause boundary identification for German, please refer to (Sidarenka et al., 2015).

sent num	verb id(s)	VC	main verb	fin	tense	mood	voice	neg	coord
1	6,7	has climbed	climbed	yes	presPerf	indicative	active	no	no
2	4,5	has crossed	crossed	yes	presPerf	indicative	active	no	no
2	13,14	can 't increase	increase	yes	present	indicative	active	yes	no

Table 4: TSV output of the annotation tool for two English sentences: "Since then, the index has climbed above 10,000. Now that gold has crossed the magic $1,000 barrier, why can't it increase ten-fold, too?"

snapshot of the tool's performance. The findings of this analysis will lead to improvement of the rules' precision in future development iterations.

6 Conclusion

We have presented an automatic tool which annotates English, French and German verbal complexes with tense, mood and voice. Our tool compensates for the lack of annotated data on this subject. It allows for large-scale studies of verbal tenses and their use within and across the three languages. This includes for instance typological studies of the temporal interpretation of tenses, or discourse studies interested in the referential properties of tense. Large-scale annotated data with reliable accuracy also creates the possibility to train classifiers, machine translation systems and other NLP tools. The same approach for extracting tense, aspect and mood could also be implemented for other languages.

Acknowledgment

This work has received funding from the DFG grant Models of Morphosyntax for Statistical Machine Translation (Phase 2), the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644402 (*HimL*), and from the European Research Council (ERC) under grant agreement No. 640550.

We thank André Blessing for developing the demo version of the tool.

References

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on EMNLP*. Jeju Island, Korea.

Francisco Costa and António Branco. 2012. Aspectual type and temporal relation classification. In *Proceedings of the 13th Conference of the EACL*. Avignon, France.

Ashwini Deo. 2012. Morphology. In Robert I. Binnick, editor, *The Oxford Handbook of Tense and Aspect*, OUP.

Annemarie Friedrich and Manfred Pinkal. 2015. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the 2015 Conference on EMNLP*. Lisbon, Portugal.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*. Phuket, Thailand.

Sharid Loáiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French verb phrase alignment in Europarl for tense translation modeling. In *Proceedings of the 9th International Conference on LREC*. Reykjavik, Iceland.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31(1):71–106.

Anita Ramm and Alexander Fraser. 2016. Modeling verbal inflection for English to German SMT. In *Proceedings of the the First Conference on Machine Translation (WMT)*. Berlin, Germany.

Roi Reichart and Ari Rappoport. 2010. Tense sense disambiguation: a new syntactic polysemy task. In *Proceedings of the 2010 Conference on EMNLP*. Massachusetts, USA.

Diana Santos. 2004. *Translation-based corpus studies Contrasting English and Portuguese tense and aspect systems*. Rodopi.

Silke Scheible, Sabine Schulte im Walde, Marion Weller, and Max Kisselew. 2013. A compact but linguistically detailed database for german verb subcategorisation relying on dependency parses from a web corpus: Tool, guidelines and resource. In *In Proceedings of the WAC-8*. Lancaster, UK.

Uladzimir Sidarenka, Andreas Peldszus, and Manfred Stede. 2015. Discourse segmentation of German texts. *Journal for Language Technology and Computational Linguistics* 30(1):71–98.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-Events, and Temporal Relations. In *Proceedings of the SemEval 2013*. Atlanta, Georgia.