ISA-17

17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation

Workshop Proceedings

Harry Bunt, editor

June 16 - 17, 2021

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-954085-20-6

Message from the General Chair

Welcome the proceedings of the online ISA workshop at IWCS 2021!

Last year, the ISA-16 workshop at LREC 2020 had to be canceled altogether, but following the decision of the LREC organisers we did publish the accepted submissions to the workshop in the proceedings that can be found online at the ISA-16 website and in the ACL anthology.

This year we are in a slightly better shape since the IWCS 2021 conference that hosts the ISA-17 workshop was planned from the beginning to be held in online form. While online presentation of papers and discussion tend to suffer from the online format, this does feel like a step forward compared to last year. In particular, in 2020 we had planned to organise two exciting shared tasks, one on the annotation of quantification phenomena and one on the representation of visual information, which were postponed to this year and will go ahead this time. The discussion notes and annotations that were submitted for these shared tasks have not been included in these proceedings, but are available at the ISA-17 website (https://sigsem.uvt.nl/isa17).

We thank the members of the ISA-17 program committee for reviewing the submitted papers timely and thoroughly, and we thank the authors of accepted papers for revising their contributions according to the original time schedule, taking the review comments into account. We thank the participants in the two shared tasks for their contributions, which promise to be most valuable for the further development of adequate semantic annotation and representation schemes. Thank you!

Harry Bunt

Organizing Committee

Harry Bunt, Tilburg University (Netherlands)
Nancy Ide, Vassar College, Poughkeepsie, NY (USA)
Kiyong Lee, Korea University, Seoul (South Korea)
Volha Petukhova, Saarland University, Saarbrücken (Germany)
James Pustejovsky, Brandeis University, Waltham, MA (USA)
Laurent Romary, INRIA/Humboldt University, Berlin (Germany)
Ielka van der Sluis, University of Groningen (Netherlands)

Program Committee

Jan Alexandersson
Johan Bos
Harry Bunt
Nicoletta Calzolari
Jae-Woong Choe
Robin Cooper
Ludivine Crible
David DeVault
Simon Dobnik
Jens Edlund
Alex Fang
Robert Gaizauskas
Koiti Hasida
Nancy Ide
Elisabetta Jezek
Nikhil Krishnaswamy
Kiyong Lee
Paul McKevitt
Adam Meyers
Philippe Muller
Rainer Osswald
Volha Petukhova
Massimo Poesio
Andrei Popescu-Belis
Laurent Preévot
James Pustejovsky
Livio Robaldo
Laurent Romary
Ielka van der Sluis
Manfred Stede
Matthew Stone
Thorsten Trippel
Carl Vogel
Menno van Zaanen
Annie Zaanen
Heike Zinsmeister

Table of Contents

Workshop Program

Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus
Purificação Silvano, António Leal, Fátima Silva, Inês Cantante, Fatima Oliveira and Alípio Mario Jorge

Towards the ISO 24617-2-compliant Typology of Metacognitive Events
Volha Petukhova and Hafiza Erum Manzoor

Annotating Quantified Phenomena in Complex Sentence Structures Using the Example of Generalising Statements in Literary Texts
Tillmann Dönicke, Luisa Gödeke and Hanna Varachkina

The ISA-17 Quantification Challenge: Background and introduction
Harry Bunt

Discourse-based Argument Segmentation and Annotation
Ekaterina Saveleva, Volha Petukhova, Marius Mosbach and Dietrich Klakow

Converting Multilayer Glosses into Semantic and Pragmatic forms with GENLIS
Rodolfo Delmonte, Serena Trolvi and Francesco Stiffoni

Unleashing annotations with TextAnnotator: Multimedia, multi-perspective document views for ubiquitous annotation
Giuseppe Abrami, Alexander Henlein, Andy Lücking, Attila Kett, Pascal Adeberg and Alexander Mehler

Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus

Purificação Silvano[1], António Leal[2], Fátima Silva[3], Inês Cantante[4], Fátima Oliveira[5] & Alípio Mário Jorge[6]

[1,2,3,4,5]University of Porto/ Centre of Linguistics [6]University of Porto/ INESC

[1]msilvano@letras.up.pt, [2]jleal@letras.up.pt,[3]mhenri@letras.up.pt, [4]cantante.ines@gmail.com, [5]moliv@letras.up.pt, [6]amjorge@fc.up.pt

Abstract

In this paper, we describe the process of developing a multilayer semantic annotation scheme designed for extracting information from a European Portuguese corpus of news articles, at three levels, temporal, referential and semantic role labelling. The novelty of this scheme is the harmonization of parts 1, 4 and 9 of the ISO 24617 *Language resource management - Semantic annotation framework.* This annotation framework includes a set of entity structures (participants, events, times) and a set of links (temporal, aspectual, subordination, objectal and semantic roles) with several tags and attribute values that ensure adequate semantic and visual representations of news stories.

1 Introduction

The development of an annotation framework can be an overwhelming task, even more when its purpose is to account for different linguistic phenomena. However, as challenging as it may be, designing an annotation scheme is an indispensable step to generate language resources that can be the starting point of fundamental corpus-based linguistic research.

When deciding on an annotation framework, one has to take into consideration several factors (Pustejovsky et al., 2017), such as main objectives of the annotation, the linguistic phenomena under analysis, the corpus genre, and the nature of the annotation, and weigh in the advantages and disadvantages of adapting/ adopting an existing model, or of creating one. Ideally, the model is custom designed to deal with all the specificities of a particular project, but also broad enough so that it can be applied to other datasets. In fact, with the growth of the Semantic Web and Linguistic Linked Data (Chiarcos et al., 2020), interoperability is key to read and to interpret linguistic resources (Ide and Pustejovsky, 2010).

With all the above-mentioned provisos in mind, we developed a multilayer semantic annotation scheme by combining three standards from the *Language resource management-Semantic annotation framework*: *Part 1- Time and events* (ISO-24617-1), *Part 4- Semantic roles* (ISO-24617-4) and Part 9- *Referential annotation framework* (ISO-24617-9). In addition to promoting interoperability, our model has proven to be able to markup manually the relevant features of the genre news to generate visual representations of their narratives. Moreover, our proposal operationalizes the integration of three different standards in the same framework, which is, to the best of our knowledge, a novelty.

This multilayer semantic annotation scheme was designed to annotate a European Portuguese corpus of news articles in three different, but complementary, levels, temporal, referential and thematic, within the Text2Story project[1], which aims to extract narratives from news, represent them in intermediate data structures, and make these available to subsequent media production processes, i.e., visualizations such as message sequence charts (MSC) and knowledge graphs (KG). In this paper, we document the decision-

[1] https://text2story.inesctec.pt/

Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pages 1–13

making process about which annotation format to adopt, what adjustments to make, and how to harmonize the three layers into an integrated and wide-ranging model.

2 Background and Motivation

News may frequently assume the format of a story that reports on current events involving one or more entities in given time and place. In addition to the main event, however, news stories typically present contextual content that allows connecting it to others, explaining the circumstances and consequences of its occurrence. It may also include other complementary information that frames, comments, clarifies, or evaluates the reported events (Caswell and Dörr, 2019; Choubey et al., 2020; cf. also van Dijk, 1985; Bell, 1991). A complete story usually answers six questions: what, who, where, when, why, and how, that is, 5W1H (a.o. Bonet-Jover et al., 2021), following a top-down organization, corresponding to an inverted pyramid discourse structure (cf. Rabe 2008), in which information flows in decreasing order of importance. A news organization structure usually features a title, a lead, and the body. In many cases, the lead or introductory paragraph condenses the answers to the above six questions and is followed by complementary information (a.o. Thomson et al., 2008; Norambuena et al., 2020). Sometimes, the answer to some of the questions is distributed throughout the text (Bonet-Jover et al., 2021). Because of this organization, events frequently follow a non-chronological order, presenting a complex time structure regarding other kinds of narratives (Zahid et al., 2019). Besides, the narrative may return to previous data, as well as adding information (a.o. van Dijk, 1985; Thomson et al., 2008; Choubey et al., 2020).

Establishing the temporal sequencing of events, their participants, and interrelations is crucial to understand the news story, and ultimately to extract the narratives to be represented graphically by means of MSC (Harel and Thiagarajan, 2003) or KG (Ehrlinger and Wöß, 2016), which is our project's main objective. These visualizations by portraying the narratives more schematically can be of great interest to news agencies, for example. The more overarching and rigorous the annotation the more informative is the visualization, and, in the case of news articles, this requires featuring participants, events and times, as well the relationships between them. For these reasons, the annotation scheme that we designed encompasses three intertwined semantic layers: temporal, referential and thematic. Since our aim was to adopt a coherent and interoperable annotation scheme with these three layers, and because none of the existing proposals satisfied these requisites, we designed an annotation scheme which compatibilizes three ISO.

3 Related work

Over the last years, there has been a proliferation of multilayer corpora, that is, corpora that "contain mutually independent forms of information, which cannot be derived from one another reliably" (Zeldes, 2019: 4). These layers can be defined in an independent way and they "are explicitly analyzed using multiple, independent annotation schemes" (Zeldes, 2019: 7), or resorting to one unique scheme that integrates all the layers. In fact, an in-depth analysis of the relevant literature reveals that there are many different types of multilayer annotation schemes. In the remainder of this section, we will only present a brief overview of some of those proposals.

One of the most well accomplished and far-reaching multilayer annotation schemes is the one developed within the *Groningen Meaning Bank* (GMB) (Basile et al., 2012; Bos et al., 2017). Besides morphological and syntactic annotation, it comprises different semantic annotation levels, such as named entity recognition, temporal features, and thematic roles. The adopted semantic formalism is an extension of *Discourse Representation Theory* (Kamp and Reyle, 1993), which renders a semantic representation (*discourse representation structures*) that unifies the various layers. Another important feature of this scheme is that it was designed to analyze linguistic phenomena in texts, instead of only sentences, and it has been used quite successfully in 10,000 texts from different genres, namely news and fables. Its implementation requires a human-aided machine annotation insofar as it employs NLP software such as an automatic tagger for named entity recognition, VerbNet (Schuler, 2005) for semantic role labelling, a semantic analyzer for coreference, and then a module Boxer (Bos, 2005, 2008; Curran

et al. 2007), responsible for the overall semantic analysis, but also relies on the input of experts and general public. Although, in terms of semantic annotation, it is one of the most complete, this scheme lacks information about more referential relations. Moreover, since the temporal annotation is based on DRT-language, it does not integrate tags about lexical and contextual meaning with bearing on temporal interpretation, namely a more diversified class of events, and other link types between events.

Other multilayer annotation schemes have been developed for *Manually Annotated Sub-Corpus* (MASC) (Ide et al., 2008), *Georgetown University Multilayer Corpus* (GUM) (Zeldes and Simonson, 2016; Zeldes, 2017), *OntoNotes* (Hovy et al., 2006), for *AMALGUM* (Gessler et al., 2020), or *SenSem* (Fernández and Vázquez, 2014), just to name a few, but none of those provide a comprehensive and harmonized semantic framework suitable to handle the linguistic phenomena that we need to address.

For European Portuguese (EP), one can point out the scheme used in CINTIL DeepBank (Branco et al., 2010), which is a corpus of Portuguese news and novels that is annotated with several grammatical information (morphological, syntactic, and semantic) for each sentence. Currently, there are 32497 sentences, mainly from news, which were semi-automatically annotated with Treebank, DependencyBank, Propbank, and LogicalFormBank (with formal representations of the sentences meanings using Minimal Recursion Semantics). However, the CINTIL DeepBank's scheme does not include a level for referential annotation, nor for temporal annotation. The fact that only the sentences that the grammar can parse are included in the corpus is a downside. Additionally, though each level of annotation can be accessed separately, a unifying formalism that combines all the layers is missing.

Regarding schemes aimed exclusively at semantic annotation, some are intended to handle a specific phenomenon, resort to non-standardized markup language, and are not widely known (cf. for an overview (Gries and Berez, 2017). Moreover, the majority deals with lexical problems, such as word disambiguation, and less with compositional semantics. The scarcity of proposals within this branch of semantics can be explained by the complexity underlying the process of annotating semantically data at a sentential and textual level. This task requires not only a great amount of time, but also a wide variety and substantial number of resources. Nonetheless, semantic schemes to represent the meaning of texts are of utmost relevance to the development of different applications.

4 The Annotation scheme

4.1 The process

Building a bootstrapping annotation scheme is a very complex and time-consuming endeavor involving different phases. After the literature review, we started by defining the tags and their attributes first for the temporal layer, then for the referential level, and finally for the semantic role labelling. To create a model, we followed the MATTER (Pustejovsky and Stubs, 2012) sub-cycle, MAMA, with four steps, (1) model, (2) annotate, (3) evaluate and (4) revise. This process allowed us to identify and resolve the scheme's inconsistencies, gaps and incompatibilities, and to gradually improve it so that it could properly account for the linguistic data, and to deliver the necessary input for the visualization task. This cycle was repeated several times until we were satisfied with the model. The annotation tool that we used, BRAT (brat rapid annotation tool) (Stenetorp et al., 2012), enabled the updates of the annotation scheme without having to rebuild the whole scheme.

4.1.1 Temporal Layer

Temporal interpretation plays a crucial part in understanding how the events are organized in natural language texts. For this reason, extraction of temporal information has been receiving a lot of attention within NLP during the past few years. One approach to extract temporal features, and eventually to rebuild chronological sequences of events, is designing a suitable annotation scheme. In this field, research has started with the extraction of time expressions in message understanding conferences (MUCs) and progressed to relating events to times (eg. Filatova and Hovy, 2001; Katz and Arosio, 2001; Song et al., 2016). From the growing investment on temporal extraction, on the one hand, and from its usefulness, on the other hand, ensued not only a significant number of corpora annotated according to different schemes,

but also annotation standards. One of these standards is TimeML (Pustejovsky et al., 2003a, 2003b), based on the work of Setzer (2001), Setzer & Gaizauskas (2000a, 2000b, 2001) and Ferro et al. (2003), from which ISO-TimeML (ISO 24617-1) stemmed.

ISO-TimeML, a model grounded on linguistic approaches (eg. Reichenbach, 1947; Comrie, 1985), defines a full-fledged markup language that permits a fine-grained annotation of time expressions, events, and temporal relations between events and between events and time expressions. Its efficacy and productivity in capturing the text's temporal structure is evidenced by corpora such as TIDES Temporal Corpus (Gerber et al., 2002), TimeBank (Pustejovsky et al., 2003b), composed of news articles, or Sun et al. (2013)'s corpus with clinical narratives. Costa (2012) and Costa and Branco (2010, 2012) use TimeML to annotate for the first time a EP corpus with temporal information, TimeBankPT. This corpus, nonetheless, only comprises the translations of texts from the original TimeBank, as well as the same annotations with some adaptations required by language specificities.

Compared to the scheme employed by TimeBankPT, the temporal tagset and linkset that we subscribe follow more closely ISO-24617-1. As expected, bearing in mind the project's main aim, that is, visualization of news narratives, and the necessity of not overloading the scheme with unnecessary information, some tags and links were excluded. Thus, for the temporal layer, our scheme incorporates two tags, event and times, and three links, temporal link (TLink), aspectual link (ALink) and subordination link (SLink).

The tag event marks eventualities (Bach, 1985), represented by tensed or untensed verbs, nominalizations, adjectives, predicative constructions or prepositional complements. The combination of all the required attributes, class, part of speech, tense, aspect, verb form, mood, modality and polarity, provides the necessary information about temporal, aspectual and modal features of events. With respect to the values for each attribute, we maintained the ones established by ISO-24617-1, namely for Italian, but added in the attribute mood the value *future* to account for its modal uses, and the modality values *dever* ('must'), *poder* ('can'), *ter de* ('have to') and *ser capaz de* ('be able to').

Regarding the tag times, we adopted a very simple scheme, which meets the needs of our project. The attributes that incorporate our annotation scheme are the required ones, according to ISO-24617-1, that is, type (date, time, duration and set) and value (the specific value of the type). We have also integrated two optional attributes: temporal function with the value publication time and anchor time, which are pertinent to process time expressions common in news articles, like *hoje* 'today', *na sexta-feira* 'Friday'.

The sequencing of the events, that is, their ordering, is essential to depict the way the narrative evolves in time. ISO-24617-1 specifies the adequate manner to establish the events timeline, as well as the relations between events and time expressions by postulating TLinks, which we integrated in our scheme. In turn, the ALink, by specifying the relation between aspectual verbs and their event arguments, gives crucial input to create the visualizations of the events. The relevance of the SLink derives from the fact that the news articles frequently include contexts of subordinating relationships between events. We omitted the measuring link (MLink) because the information it conveys is already captured to a certain extent by the value duration for the attribute type of tag times. The values for the three links of our model are the ones proposed by ISO-24617-1.

4.1.2 Referential Layer

Pointing out to the referring expressions in a text, identifying the discourse entities denoted by those expressions, and establishing the links between them are key tasks to reference annotation, and underly referential phenomena in discourse, such as anaphora.

In our corpus, those referring expressions correspond to named entities, or participants that play an important role in the story. Therefore, we needed a framework to deal with named entities recognition and their relation throughout the news texts. ISO-24617-9 met these needs, as it is a meta-model of referential annotation that articulates the discourse domain with the linguistic domain, contributing to a comprehensive representation of the discourse entities, the referring expressions that denote them, and their relations.

Despite following the standard in its overall guidelines, we did not annotate all its categories, and both discourse entity structures and referential

expression structures were kept as simple as possible, to avoid overloading the process of annotation: the former include only information concerning the lexical head (noun, pronoun), whereas the latter include information concerning domain (individuation and types) and involvement. The individuation attribute, with the values set, individual and mass, follows ISO-24617-9 definitions, while for involvement we defined the values: *0* (the empty set); *1* (a set with only one entity); *>1* (a set with more than one entity, but less than the totality of entities in the domain); *all* (the totality of entities in the domain = universal quantification); *undef* (undefined involvement).

As for types, since ISO 24617-9 does not provide a typology of named entities, we selected, considering our corpus text genre and the purpose of the project, a tagset of six named entities: PER, ORG, LOC, OBJ, NAT, OTHER. In fact, the definition of named entities is neither easy nor consensual, and there are several typologies for their classification, being the number and types of entities influenced by factors, such as the domain from which they are extracted or the purpose of its classification (for a survey on this topic, see, a.o. Nouvel et al., 2016; Goyal et al., 2018). This tagset is an adaptation of general categories depicted in the named entity classification typologies used in many other corpora, including multilayer ones. The first three named entities are common to all the annotated corpora while the others may vary.

In what concerns the relations included in ISO-24617-9, we did not include in our specifications the lexical relational links between entity structures and referring expressions (eg. synonym, antonym, hyponym, meronym), the referential status of referring expressions (old/new), and the properties of discourse entities (abstractness, animacy, alienability, natural gender and cardinality), because they were not necessary for the visual representations of news. As a matter of fact, it is more useful for visualization to mark two linguistic expressions as referring to the same participant. Thus, our analysis only considers the proposed objectal links (objectalIdentity, partOf, subset, memberOf and referentialDisjunction) between discourse entities, which allows to represent nominal anaphora's mechanisms. Unlike many studies that focus on anaphora resolution and depict only coreferential mechanisms, leaving out other types of relations, the adopted framework allows for the marking of different types of anaphoric linkage between entities, namely direct and indirect anaphora.[2]

4.1.3 Semantic Role Layer

The task of semantic role labelling for English texts usually uses one of the following frameworks (see also ISO 24617-4, Annex B): FrameNet (Baker et al., 1998), VerbNet (Schuler, 2005), PropBank (Palmer et al., 2005), EngVallex (Cinková, 2006), and LIRICS (Petukhova and Bunt, 2008).

As for EP data, there are some proposals that approach the issue of semantic role labelling, typically using the methodology of PropBank and VerbNet. However, these proposals have a very narrow scope, working with small datasets and small lists of (typically) verbs. Some examples of these works are PropBankPT (Branco et al., 2012), a corpus of 3406 sentences translated from the *Wall Street Journal*, and annotated with information concerning constituency structure (phrase constituency and grammatical relations) and semantic roles; and CINTIL-PropBank (Branco et al., 2012), a corpus of 10039 sentences extracted from news and novels, and annotated with information concerning constituency structure and semantic roles. There is also ViPer (Talhadas et al., 2013), a verbal lexical database with information about the verb's arguments semantic roles (using PropBank approach) manually annotated. However, there are some aspects of the semantic roles list that is used that can be problematic for our project (for instance, event-denoting nouns are treated as arguments of the "occurrence" type, instead of being treated as events, like in ISO-24617-1).

So, the semantic role labelling task in our project could not be based on previous work done for EP, and it had to be done from scratch. The easier way to do so was to use some established framework and adapt it to EP, but the methodology typically used in frameworks designed for English (eg. FrameNet) requires that, for each verb, a frame be

[2] The Universal Anaphora initiative (https://universalanaphora.github.io/UniversalAnaphora/) has been working towards a proposal markup scheme compatible with Universal Dependencies, and that codifies different aspects of the anaphoric phenomena.

constructed, and the construction of each frame entails many examples with the same verb and their analysis (to identify all the meanings the verb can have and all the constructions in which it can occur), to determine its semantic selection. This work would be colossal, and impracticable taking into account the time frame and objectives of the project. Therefore, we needed a framework that would allow semantic annotation to be limited to the analysis of concrete examples of the news to be annotated. We started working with the framework provided by LIRICS, which was the most appropriate for the task. Furthermore, as LIRICS was the basis for the construction of the ISO standard for thematic annotation, there would be fewer potential problems when integrating semantic role annotation with referential and temporal annotation.

Consequently, in our project, we annotate semantic roles following ISO-24617-4 specifications in what concerns semantic roles. We do not construct entity structures, nor event structures in this level of annotation. Instead, we use the entity structures constructed in the referential annotation to deal with non-event discourse entities, and the entity structures constructed in the temporal annotation to deal with event discourse entities. The semantic role annotation consists in establishing the thematic relation between predicates and their arguments and modifiers.

4.2 Harmonizing Different Layers

The foregoing describes how the markup language used in each layer of our annotation scheme was extracted from three different standards. Although they comply with the principles for semantic annotation (ISO-24617-6), in fact, they were elaborated separately and assynchronically, and they lack information about how to combine them with each other. ISO-24617-6, in addition to defining some overall guidelines for the semantic annotation framework (SemAF), attempts at tackling some overlaps and inconsistencies between the different parts of the SemAF, but its coverage is limited. This means that, when combining different parts of the SemAF, as we did, it is expected that not only some incompatibilities may arise, but also some loose-ends and gaps may be left unsolved. Proposals such as Bunt (2019) improve the absence of some information in one particular part of SemAF by resorting to some notations from other parts of the ISO. Gaizauskas and Alrashid (2019), for instance, put forward a scheme with some annotations from ISO-24617-1/7, but do not refer to issues related to incompatibilities. Therefore, in the process of constructing our model, we had to overcome these difficulties in order to obtain a fully integrated scheme.

We began by modelling the types of structures as entity structures and link structures, and defined subtypes for each type, as described in Figure 1.

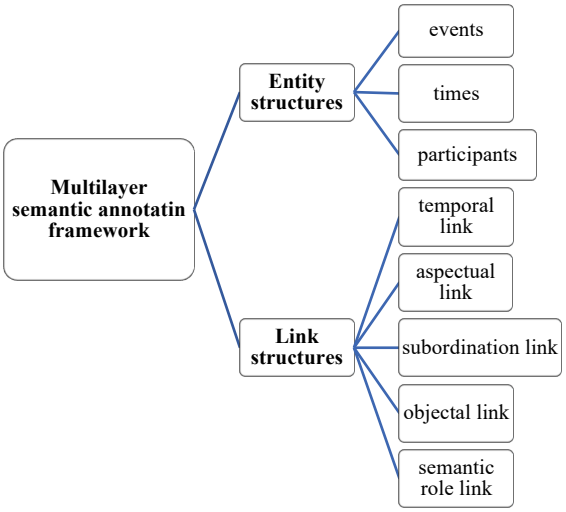


Figure 1: Text2story semantic annotation framework

This annotation structure is the first step to guarantee that all the layers are combined into a coherent annotation scheme. The entity structures, regardless of the layer to which they are associated, are available to be related among them by different types of link structures. Such unifying approach facilitates a uniform semantic representation in discourse representation structures (DRS).

The next step was to decide on the attributes and their respective values, so the information they codified would be compatible and not repetitive, as explained in the previous sections. The final annotation scheme is presented in Table 1.

ENTITY STRUCTURES		
EVENTS	class	occurrence, state, reporting, perception, aspectual, I-action, I-state
	type	state, process, transition
	pos	verb, noun, adjective, preposition
	tense	present, past, future, imperfect, none
	aspect	progressive, perfective, imperfective, imperfective-progressive, perfective-progressive, none
	vform	none, gerundive, infinitive, participle
	mood	none, subjunctive, conditional, future, imperative
	modality	*dever, poder, ter de, ser capaz de*
	polarity	negative, positive
TIME	type	date, time, duration, set
	value	specific value
	temporal function	publication_time
	anchortime	time ID (select relevant time)
PARTICIPANTS	lexical head	noun, pronoun
	domain	individuation: set, individual, mass
		types: per, org, loc, obj, nat, other
	involvement	0,1, >1, all, undefined
LINK STRUCTURES		
Temporal links		before, after, includes. is_included, during, simultaneous, identity, begins, ends, begun_by, ended_by
Aspectual links		initiates, culminates, terminates, continues, reinitiates
Subordination links		intensional, evidential, neg_evidential, factive, counter_factive, conditional
Objectal links		objectalIdentity, partof, subset, memberOf, referentialDisjunction
Semantic role links		agent, source, location, path, goal, time, theme, instrument, partner, patient, pivot, cause, beneficiary, result, reason, purpose, manner, medium, means, setting, initialLocation, finalLocation, distance, amount, attribute

Table 1: Text2story annotation scheme

The harmonization of the different annotation layers using ISO-standards presented us with some mismatches between the three ISOs, which had to be addressed and solved. As an illustration, we present two of those issues.

Concerning markables, while the thematic annotation specifications in ISO 24617-4 foretold that a clause may receive a semantic role, the referential ISO does not stipulate any entity structure for clauses. Our solution to this problem was to mark the event structure corresponding to the verbal predicate of the subordinated clause so that the semantic role link can be set up. Accordingly, in a sentence like *John said that Mary went to Porto* the chunk that is linked to "said" by the semantic role theme is not the whole clause, but only the verb "went", because it has been already associated to an entity structure, more precisely to an event structure, in the temporal layer, contrary to the clause. This solution adopts a Neo-Davidsonian perspective of the relation between events and their arguments and considers that all entities with an event structure annotated in the temporal level correspond to an event argument of a predicate. So, in a Neo-Davidsonian version, the sentence above would have the following logical form: $\exists e^1$ [SAY (e^1) & AGENT (e^1, John) & $\exists e^2$ [GO ($\exists e^2$) & AGENT ($\exists e^2$, Mary) & TO ($\exists e^2$, Porto) & THEME (e^1, e^2)]].

However, some problems are of more difficult resolution. ISO-24617-4 envisages that some adverbial phrases may be attributed the semantic role of manner, like "tightly" in the sentence *The tiny stick was fastened tightly to his wrist* (ISO-24617: 23). Nonetheless, "tightly" in our framework (and in the relevant ISO-standards, for that matter) cannot be marked as any kind of entity structure. We could simply disregard it because it is a modifier, but in some cases manner adverbial phrases are complements (*The child behaved badly*), conveying pertinent information to the story, and, hence, they should be annotated. At this moment, we still have no means to come to grips with this conundrum.

Despite the above-mentioned hurdles, we have been able to conciliate three ISO-standards and produce a consistent and complete multilayer semantic annotation scheme, which not only adequately serves the purpose of our project, but may also contribute to other annotations' schemes.

5 An Annotated Example

In our model, the annotation procedure consists of three stages. Example (1) will serve to illustrate the three stages.

(1) 20/03/2021
Cientistas que estudavam a erupção de um vulcão da Islândia decidiram esta sexta-feira usar a lava expelida da cratera para assar salsichas.
Scientists that were studying the eruption of a volcano of Iceland decided this Friday to use the lava expelled from the crater to roast sausages.

In the first stage, the annotator marks the entity structures of events and times, and, then, the temporal, aspectual and subordination links are established.

EVENTS
e1=*estudavam* class=occurrence type=process pos=verb tense=past aspect=imperfective polarity=pos vform=none mood=none
e2=*erupção* class=occurrence type=process pos=noun tense= none aspect= none polarity= pos vform=none mood=none
e3=*decidiram* class=occurrence type=transition pos=verb tense= past aspect=perfective polarity= pos vform= none mood= none
e4= *usar* class=occurrence type= process pos=verb tense=none aspect=none polarity=pos vform=infinitive mood= none
e5=*expelida* class=occurrence type=transition pos=verb tense=past aspect= perfective polarity= pos vform=participle mood=none
e6=*assar* class=occurrence type=process pos=verb tense=none aspect=none polarity=pos vform= infinitive mood=none

TIME EXPRESSIONS
t1=*20/03/2021* type=date value=20-03-2021 FunctionInDocument= publication time
t2=*esta sexta-feira* type=date value=19-03-2021 AnchorTimeID=t1

TLINK
e2 before e1
e3 is_included e1
e3 is_included t1
e4 after e3
e5 before e3
e6 simultaneous e4

SLINK
e4 intensional e3
e6 intensional e4

In the second stage, the participants are identified, and they are related to each other by objectal links.

PARTICIPANTS
p1=*cientistas que estudavam a erupção de um vulcão na Islândia* lexical head=noun individuation=individual type =per involvement= >1
p2=*que* head=pronoun individuation=individual type =per involvement= >1
p3=*um vulcão da Islândia* head=noun individuation=individual type =per involvement=1
p4=*a lava expelida da cratera* head=noun individuation= mass type =nat involvement=1
p5= *a lava* head=noun individuation=mass type=nat involvement= undef
p6=*a cratera* head=noun individuation= individual type =nat involvement=1
p7=*salsichas* head=noun individuation= individual type =obj involvement=>1

OBJECTAL LINKS
p2 ObjIdentity p1
p5 partOf p3
p6 partOf p3

In the third stage, the annotator connects participants to events by semantic role links.

SEM_ROLE_LINK
p1=agent (e3)
p2=agent (e1)
p3=patient (e2)
p4=instrument (e4)
p5=theme (e5)
p6=initial location (e5)
p7=patient (e6)
e6=purpose (e4)
p1=agent (e4)
p1=agent (e6)
e2=theme (e1)
e4=theme (e3)

After carrying out this manual annotation in the annotation tool BRAT [3], our project's pipeline includes two more modules: the Brat2DRS, which takes the annotation file generated by Brat, parses it, and creates a DRS representation; and the BRAT2Viz, which takes as input the DRS representation, and deploys a web application that produces the visualizations in the form of MSC or KG (Amorim et al., 2021).

6 Conclusion

In this paper, we present an annotation framework for news articles in EP that aims to provide the input for visualization processes. First, we determined what type of information was necessary to account for events and participants in the narratives, and decided that three annotation layers - temporal, referential and thematic - were required. The next step was to decide which tags and links should be used in each layer to fulfill the annotation purposes. Since interoperability is crucial when we talk about semantic resources, three standards ISO 24617-1/4/9 were utilized to create a multilayer semantic annotation scheme. Notwithstanding the fact that these standards are, in fact, themselves three parts of the same standard, when combined, some inconsistencies arise. So, we had to harmonize the three layers, to attain a cohesive annotation framework. Additionally, we sought to balance the amount of information needed to capture the news stories and the load of the annotation process.

Although this model was built to capture the structure of stories in news in EP, its scope is not limited to news nor to EP, as it can be extended to other narrative texts and other languages with some adaptations to deal with genre and language specificities. Moreover, the integration of three different layers in a single annotation framework enables formal semantic representation with DRS, which acts as an intermediate language to generate visualizations in the form of knowledge graphs, for instance.

In the future, we intend to endow our annotation scheme with more granularity. To this end, ISO standard for spatial information (ISO 24617-7) will be added to our framework. For now, spatial annotation has relied on the tags, attributes and links available in the referential and thematic layers. Likewise, a more detailed information regarding quantification of participants and of events is a component to be improved in the future. At this moment, this kind of information has a very simplified representation solely in the referential layer, which does not fully represent the different possibilities of quantification over entities.

[3] https://nabu.dcc.fc.up.pt/brat/#/examples_demos/paper_ISA-17

Acknowledgments

The authors wish to thank the reviewers for their constructive comments. This research is financed by the ERDF – European Regional Development Fund through the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project PTDC/CCI-COM/31857/2017 (NORTE-01-0145-FEDER-03185). The usual disclaimers apply.

References

Amorim, Evelin; Ribeiro, Alexandre; Cantante, Inês; Jorge, Alípio; Santana, Brenda; Nunes, Sérgio; Silvano, Purificação; Leal, António; & Campos, Ricardo (2021). Brat2Viz: a Tool and Pipeline for Visualizing Narratives from Annotated Texts. In *Text2Story 2021. Fourth International Workshop on Narrative Extraction from Texts*. (pp. 49-56). Lucca, Italy: CEUR Workshop Proceedings, CEUR-WS.org.

Bach, Emmon (1985). The algebra of events. *Linguistics and Philosophy*, *9*, 5–16.

Baker, Collin; Fillmore, Charles; & Lowe, John (1998). The Berkeley FrameNet project. In *Proceedings of the Conference on 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics.* (pp. 86–90). Montréal, Quebec: Université de Montréal. Retrieved from https://www.aclweb.org/anthology/P98-1013/

Basile, Valerio; Bos, Johan; Evang, Kilian; & Venhuizen, Noortje J. (2012). Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. (pp. 3196–3200). Istanbul, Turkey: ELRA. Retrieved from https://www.aclweb.org/anthology/L12-1299/

Branco, António; Costa, Francisco; Silva, João; Silveira, Sara; Castro, Sérgio; Avelãs, Mariana; Pinto, Clara; & Graça, João (2010). Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: the CINTIL DeepGramBank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 1810–1815). Valletta, Malta: ELRA. Retrieved from http://www.di.fc.ul.pt/~ahb/pubs/2010BrancoCostaSilvaEtAl.pdf

Branco, António; Carvalheiro, Catarina; Pereira, Sílvia; Avelãs, Mariana; Pinto, Clara; Silveira, Sara; Costa, Francisco; Silva, João; Castro, Sérgio; & Graça, João (2012). A PropBank for Portuguese: The CINTIL-PropBank. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* (pp. 1516–1521). Istanbul, Turkey: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/summaries/373.html

Bell, Allan (1991). *The Language of News Media*. Oxford: Blackwell.

Bonet-Jover, Alba; Piad-Morffis, Alejandro; Saquete, Estela; Martínez-Barco, Patricio; & García-Cumbreras, Miguel Ángel (2021). Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. *Expert Systems with Applications, 169,* 1–19. doi: 10.1016/j.eswa.2020.114340

Bos, Johan (2005). Towards wide-coverage semantic interpretation. In *Proceedings of IWCS-6*. (pp. 42–53). Tilburg, The Netherlands. Retrieved from https://www.let.rug.nl/bos/pubs/Bos2005IWCS.pdf

Bos, Johan (2008). Wide-Coverage Semantic Analysis with Boxer. In Johan Bos & Rodolfo Delmonte (eds.). *Semantics in Text Processing. STEP 2008 Conference Proceedings, volume 1 of Research in Computational Semantics*. (pp. 277–286). College Publications.

Bos, Johan; Basile, Valerio; Evang, Kilian; Venhuizen, Noortje J.; & Bjerva, Johannes (2017). The Groningen Meaning Bank. In Nancy Ide & James Pustejovsky (eds.), *Handbook of Linguistic Annotation* (pp. 463–496). USA: Springer. ISBN 978-94-024-0879-9. doi.org/10.1007/978-94-024-0881-2_18

Bunt, Harry (2019). *Plug-ins for content annotation of dialogue acts* . In *Proceedings of the 15th Joint ACL - ISO Workshop on Interoperable Semantic Annotation* (ISA-15) (pp.33–45). Gothenburg, Sweden. Retrieved from https://sigsem.uvt.nl/isa15/ISA-15_proceedings.pdf

Caswell, David; & Dörr, Konstantin (2019). Automating Complex News Stories by Capturing News Events as Data. *Journalism Practice, 13(8),* 951–955. doi.org/10.1080/17512786.2019.1643251

Chiarcos, Christian; Klimek, Bettina; Fäth, Christian; Declerck, Thierry; & McCrae, John P. (2020). On the Linguistic Linked Open Data Infrastructure. In *Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020)* (pp. 8–15). Language Resources and Evaluation Conference (LREC). Marseille, France. Retrieved from https://www.aclweb.org/anthology/2020.iwltp-1.2/

Choubey, Prafulla Kumar; Lee, Aron; Huang, Ruihong; & Wang, Lu (2020). Discourse as a Function of Event: Profiling Discourse Structure in News. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. (pp. 5374–5386). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.478/

Cinková, Silvie (2006). From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (pp. 2170–2175). Genova, Italy: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L06-1058/

Comrie, Bernard (1985). *Tense*. Cambridge: Cambridge University Press.

Costa, Francisco (2012). *Processing Temporal Information in Unstructured Documents*. (Doctoral dissertation, Universidade de Lisboa). Retrieved from https://repositorio.ul.pt/handle/10451/8639

Costa, Francisco; & Branco, António (2010). Temporal information processing of a new language: Fast porting with minimal resources. In *ACL2010—Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 671–677). Uppsala, Sweden. Retrieved from https://www.aclweb.org/anthology/P10-1069/

Costa, Francisco; & Branco, António (2012). Extracting temporal information from Portuguese texts. In Helena Caseli; Aline. Villavicencio; António Teixeira; & Fernando Perdigão (Eds). *Computational Processing of the Portuguese Language. PROPOR 2012. Lecture Notes in Computer Science, vol 7243* (pp. 99–105). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-28885-2_11

Curran, James; Clark, Stephen; & Bos, Johan (2007). Linguistically Motivated Large-Scale NLP with CandC and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 33–36). Prague, Czech Republic. Retrieved from https://www.aclweb.org/anthology/P07-2009/

Ehrlinger, Lisa; & Wöß, Wolfram (2016). Towards a definition of knowledge graphs. In: SEMANTiCS (Posters, Demos, SuCCESS), 48, 1-4. http://ceur-ws.org/Vol-1695/paper4.pdf

Fernández-Montraveta, Ana; & Vázquez, Gloria (2014). The SenSem Corpus: an annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, *10 (2),* 273–288. doi.org/10.1515/cllt-2013-0026

Ferro, Lisa; Gerber, Laurie; Mani, Inderjeet; & Wilson, George (2003). *TIDES 2003 standard for the annotation of temporal expressions* (technical report). The MITRE Corporation. Retrieved from https://www.mitre.org/sites/default/files/pdf/ferro_tides.pdf

Filatova, Elena; & Hovy, Eduard (2001). Assigning Time-Stamps to Event-Clauses. In *Proceedings of the ACL-EACL 2001 Workshop on Temporal and Spatial Information Processing* (pp. 88–95). Toulouse: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W01-1313/

Gaizauskas, Robert, & Alrashid, Tarfah. (2019) SceneML: A Proposal for Annotating Scenes in Narrative Text, In *Proceedings of the 15th Joint ACL - ISO Workshop on Interoperable Semantic Annotation* (ISA-15) (pp.13–21), Gothenburg, Sweden. Retrieved from https://sigsem.uvt.nl/isa15/ISA-15_proceedings.pdf

Gerber, Laurie; Ferro, Lisa; Mani, Inderjeet; Sundheim, Beth; Wilson, George; & Kozierok, Robyn (2002). Annotating Temporal Information: From Theory to Practice. In *Proceedings of the 2nd international conference on Human Language Technology Research* (pp. 226–230). San Francisco, CA: Morgan Kaufmann Publishers. Retrieved from https://dl.acm.org/doi/10.5555/1289189.1289202

Gessler, Luke; Peng, Siyao Logan; Liu, Yang; Zhu, Yilun; Behzad, Shabnam; & Zeldes, Amir (2020). AMALGUM - A free, balanced, multilayer English web corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* (pp. 5267–5275). Marseille: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.648/

Goyal, Archana; Vishal, Gupta; & Kumar, Manish (2018). Recent Named Entity Recognition and Classification techniques: A systematic review. *Computer Science Review, 29,* 21–43. doi.org/10.1016/j.cosrev.2018.06.001

Gries, Stefan Th.; & Berez, Andrea L. (2017). Linguistic Annotation in for Corpus Linguistics. The Groningen Meaning Bank. In Nancy Ide & James Pustejovsky (Eds.). *Handbook of Linguistic Annotation* (pp. 379–410). USA: Springer. ISBN 978-94-024-0879-9.

Harel, David; & Thiagarajan, P.S. (2003). Message Sequence Charts. In Luciano Lavagno; Martin Grant; & Bran Selic (Eds.). *UML for Real: Design of Embedded Real-Time Systems* (pp. 77–105). USA: Springer. ISBN 978-0-306-48738-5. https://doi.org/10.1007/0-306-48738-1_4

Hovy, Eduard; Marcus, Mitchell; Palmer, Martha; Ramshaw, Lance; & Weischedel, Ralph (2006) OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers.* (pp. 57–60). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N06-2015/

Ide, Nancy; Baker, Collin; Fellbaum, Christiane; Fillmore, Charles; & Passonneau, Rebecca (2008). MASC: The manually annotated Sub-Corpus of American English. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008* (pp. 2455-2460). European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/617_paper.pdf

Ide, Nancy; & Pustejovsky, James (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proc. of the 2nd International Conference on Global Interoperability for Language Resources (ICGL)*. Hong Kong, China. Retrieved from https://www.cs.vassar.edu/~ide/papers/ICGL10.pdf

ISO24617-1:2012, Language resource management- Semantic annotation framework (SemAF) - Part 1: Time and events (SemAF-Time, ISO-TimeML)

ISO-24617-4: 2014, Language resource management- Semantic annotation framework (SemAF) - Part 4: Semantic roles (SemAF-SR)

ISO 24617-6: 2016, Language resource management- Semantic annotation framework (SemAF) - Part 6:

Principles of semantic annotation (SemAF Principles)

ISO 24617-7: 2019, Language resource management-Spatial information (SemAF) - Part 7: Reference annotation framework (ISO-Space)

ISO 24617-9: 2019, Language resource management-Semantic annotation framework (SemAF) - Part 9: Reference annotation framework (RAF)

Kamp, Hans; & Uwe Reyle (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer Academic Publishers.

Katz, Graham; & Arosio, Fabrizio (2001). The Annotation of Temporal Information in Natural Language Sentences. In *Proceedings of ACL-EACL 2001, Workshop for Temporal and Spatial Information Processing* (pp. 104–111). Association for Computational Linguistics. Toulouse. Retrieved from https://www.aclweb.org/anthology/W01-1315/

Norambuena, Brian Keith; Horning, Michael; & Mitra, Tanushree (2020). Evaluating the Inverted Pyramid Structure through Automatic 5W1H Extraction and Summarization. *Computation Journalism Symposium,* 1–7. Retrieved from https://par.nsf.gov/biblio/10168974

Nouvel, Damien; Ehrmann, Maud; & Rosset, Sophie (2016). *Named Entities for Computational Linguistics.* ISTE/Wiley, UK/USA.

Palmer, Martha; Gildea, Daniel; & Kingsbury, Paul (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistis, 31 (1),* 71–106. Retrieved from https://www.aclweb.org/anthology/J05-1004/

Petukhova, Volha; & Bunt, Harry (2008). LIRICS semantic role annotation: design and evaluation of a set of data categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 39–45). Marrakech, Morocco: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L08-1428/

Pustejovsky, James; Bunt, Harry; & Zaenen, Annie (2017). Designing Annotation Schemes: From Theory to Model. The Groningen Meaning Bank. In Nancy Ide & James Pustejovsky (Eds.). *Handbook of Linguistic Annotation* (pp. 21–72). USA: Springer. ISBN 978-94-024-0879-9.

Pustejovsky, James; & Stubbs, Amber (2012). *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., USA.

Pustejovsky, James; Castaño, José; Ingria, Robert; Saurí, Roser; Gaizauskas, Robert; Setzer, Andrea; & Katz, Graham (2003a). TimeML: robust specification of event and temporal expressions in text. In *IWCS-5, Fifth International Workshop on Computational Semantics* (pp. 28–34). Retrieved from https://www.aaai.org/Papers/Symposia/Spring/2003/SS-03-07/SS03-07-005.pdf

Pustejovsky, James; Hans, Patrick; Saurí, Roser; See, Andrew; Gaizauskas, Robert; Setzer, Andrea; Radev, Dragomir; Sundheim, Beth; Day, David; Ferro, Lisa; & Lazo, Marcia (2003b). The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics, Lancaster* (pp. 647–656). Retrieved from https://www.researchgate.net/publication/228559081_The_TimeBank_corpus

Rabe, Robert (2008). Inverted Pyramid. In Stephen L. Vaughn (Ed.). *Encyclopedia of American Journalism*. (pp. 223–225). New York: Routledge.

Reichenbach, Hans (1947). *Elements of Symbolic Logic*. New York: Macmillan.

Schuler, Karin (2005). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon* (Doctoral dissertation, University of Pennsylvania). Retrieved from https://verbs.colorado.edu/~kipper/Papers/dissertation.pdf

Setzer, Andrea (2001). *Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study* (Doctoral dissertation, University of Sheffield). Retrieved from http://etheses.whiterose.ac.uk/14436/

Setzer, Andrea; & Gaizauskas, Robert (2000a). Annotating events and temporal information in newswire text. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)* (pp. 1287–1293). Athens, Greece: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L00-1241/

Setzer, Andrea; & Gaizauskas, Robert (2000b). Building a temporally annotated corpus for information extraction. In *Proceedings of the Information Extraction Meets Corpus Linguistics Workshop at the 2nd International Conference on Language Resources and Evaluation (LREC 2000)* (pp. 9–14). Athens, Greece: European Language Resources Association (ELRA). Retrieved from http://staffwww.dcs.shef.ac.uk/people/R.Gaizauskas/research/papers/lrec00-ie-meets-cl-ter.pdf

Setzer, Andrea; & Gaizauskas, Robert (2001). A pilot study on annotating temporal relations in text. In *ACL 2001 Workshop on Temporal and Spatial Information Processing* (pp. 73–80). Toulouse, France. Retrieved from https://www.aclweb.org/anthology/W01-1311.pdf

Song, Zhiyi; Bies, Ann; Strassel, Stephanie; Ellis, Joe; Mitamura, Teruko; Dang, Hoa Trang; Yamakawa, Yukari; & Holm, Sue (2016). Event Nugget and Event Coreference Annotation. In *Proceedings of the Fourth Workshop on Events*. (pp. 37–45). San Diego, CA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W16-1005/

Stenetorp, Pontus; Pyysalo, Sampo; Topić, Goran: Ohta, Tomoko; Ananiadou, Sophia; & Tsujii, Juníchi (2012). BRAT: a web-based tool for NLP-assisted text annotation. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. (pp. 102–107). Avignon, France: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/E12-2021.pdf

Sun, Weiyi; Rumshisky, Anna; & Uzuner, Ozlem (2013). Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics, 46, Supplement,* 5–12. doi: 10.1016/j.jbi.2013.07.004

Talhadas, Rui; Mamede, Nuno; & Baptista, Jorge (2013). Semantic Roles for Portuguese Verbs. In *32nd International Conference on Lexis and Grammar* (pp. 127–132). Faro: Universidade do Algarve. Retrieved from https://www.inesc-id.pt/ficheiros/publicacoes/11312.pdf

Thomson, Elizabeth A.; White, Peter R.R; & Kitley, Philip (2008). "Objectivity" and "hard news" reporting across cultures: comparing the news report in English, French, Japanese and Indonesian journalism. *Journalism Studies*, *9(2),* 212–228. doi.org/10.1080/14616700701848261

Van Dijk, Teun A. (1985). Structures of news in the press. In Teun A. Van Dijk (Ed). *Discourse and Communication: New Approaches to the Analysis of Mass Media Discourse and Communication*. (pp. 69–93). Berlin/New York: Walter de Gruyter.

Zahid, Iqra; Zhang, Hao; Boons, Frank; & Batista-Navarro, Riza (2019). Towards the Automatic Analysis of the Structure of News Stories. In Alípio Jorge; Ricardo Campos; Adam Jatowt & Sumit Bhatia (Eds.). *Proceedings of the Text2StoryIR'19 Workshop* (pp. 71–79). Cologne, Germany. Retrieved from http://ceur-ws.org/Vol-2342/paper9.pdf

Zeldes, Amir (2017). The GUM Corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation, 51(3),* 581–612. doi: 10.1007/s10579-016-9343-x

Zeldes, Amir (2019). Multilayer Corpus Studies. New York and London: Routledge.

Zeldes, Amir; & Simonson, Dan (2016). Different flavors of GUM: Evaluating genre and sentence type effects on multilayer corpus annotation quality. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW X 2016)* (pp. 68–78). Berlin: Association for Computational Linguistics. doi: 10.18653/v1/W16-1709

Towards the ISO 24617-2-compliant Typology of Metacognitive Events

Volha Petukhova and **Hafiza Erum Manzoor**
Spoken Language Systems Group, Saarland Informatics Campus
Saarland University, Saarbrücken, Germany
{v.petukhova,hemanzoor}@lsv.uni-saaarland.de

Abstract

The paper presents ongoing efforts in design of a typology of metacognitive events observed in a multimodal dialogue. The typology will serve as a tool to identify relations between participants' dispositions, dialogue actions and metacognitive indicators. It will be used to support an assessment of metacognitive knowledge, experiences and strategies of dialogue participants. Based on the multidimensional dialogue model defined within the framework of Dynamic Interpretation Theory and ISO 24617-2 annotation standard, the proposed approach provides a systematic analysis of metacognitive events in terms of dialogue acts, i.e. concepts that dialogue research community is used to operate on in dialogue modelling and system design tasks.

1 Introduction

Daily life is replete with determinations about the reliability of our own thoughts and feelings as well as attributions about the thoughts and feelings of others. These metacognitive capacities underlie cognitive and social adaptation, influence decision-making, can enhance self-efficacy. Metacognition enables cognitive control needed for people to anticipate the future task demands, improves learning and performance on complex memory tasks, knowledge transfer and task switching (Taatgen, 2013). Cognitive models of metacognitive processes, when integrated into human-computer dialogue, transform the dialogue system from a reactive dialogue participant into a proactive learner, accomplished multi-tasking planner and adaptive decision maker (Malchanau et al., 2018).

Metacognitive capabilities of existing interactive systems, even of complex smart learning environments (Spector, 2014), are still rather limited so as metacognitive strategies used. To exploit the full potential of efficient metacognitive support in a dialogue system, big multimodal data samples are required to reliably identify metacognitive states accounting for a complexity of multidimensional contingencies between tasks, performed actions and participants' cognitive and emotional dispositions. An elaborate computational model of (meta)cognitive states calls, in the first place, for a typology of *metacognitive events* – reflexive activities that express the sender's mindful awareness of own and others cognitive processes, e.g. checking out and verification of attention, recognition, understanding, evaluation and regulation of content, thought processes, attitudes, preferences, assumptions and emotions. Metacognitive events should be computable/learned from a range of low level multi-sensory and psycho-physiological indicators (markers). Methods are required to transform multimodal data in a meaningful way to enable appropriate measurements of metacognition, adaptive decision-making and efficient coordination of multiple dialogue tasks. The main goal of the presented study is to provide a theoretical framework, methodological insights and experimental design to model relevant metacognitive processes, and specify a set of recognizable and measurable indicators to assess metacognition in dialogue.

The paper is structured as follows. Section 2 reviews methods to assess metacognition in interactive setting. In Section 3, we specify the model of metacognitive processes within the framework of Dynamic Interpretation Theory (DIT). We adapt the established metacognition assessment instruments in order to discover potential correlations between dialogue acts and metacognitive events. Section 4 presents experimental design featuring data collection, processing and ISO 24617-2 compliant annotation protocols. We wrap up the paper by outlining expected project outcomes.

2 Metacognition Assessment Instruments

Assessment of metacognition traditionally involves **self-reported** measurements. The most widely used Metacognition Questionnaire (MCQ, Cartwright-Hatton and Wells (1997)) evaluates fac-

Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pages 14–19
June 16–17, 2021.

tors related to positive and negative metacognitive beliefs, metacognitive monitoring and judgements of cognitive confidence. Questionnaires are however of limited value since they are subjective and not always accurate (Schraw, 2009).

There are two online methods proposed to assess metacognition: *thinking aloud* and *reflection when prompting*. Participants speak about their own cognitive states or processes and their understanding of partner's states and processes, or are prompted to reflect on the reasons why they chose specific actions – **verbalized metacognition**. The methods enable assessment of three elements of metacognition - experiences (e.g. confidence, confusion), knowledge (e.g. gaps), and strategies (e.g. actions). Think-aloud and prompting protocols provide rich information about the metacognitive processes when performing a task and are powerful predictors of test performance (Bannert and Mengelkamp, 2008). Verbalization methods are proven valid, but time consuming. Moreover, elicitation of explicit monitoring, reflection and regulation moments may disrupt or even break down the interaction process, distort its naturalness, trigger attention theft, increase cognitive load and impact negatively participants' engagement.

There is research performed on the **psychophysiological** measurement of metacognition. Physiological measures make use of EEG electroencephalography (Wokke et al., 2020), heart rate (Meessen et al., 2018), and pupil dilation (Lempert et al., 2015), but require rather complex and often expensive hard- and software set ups. Other methods exploit information about interlocutor's behaviour via **log files** and efficiently combine it with questionnaire data (Linek et al., 2008).

Recently, increasing computational power and technological advances opened up new data-driven assessment scenarios. A huge diversity of inexpensive tracking and sensing devices enable rather exhaustive **real-time monitoring** and immediate assessment of affective cognitive states, including metacognitive aspects (Gašević et al., 2015). Significant progress has been booked in automatic affective cognitive state recognition from speech and visual signals (Kapoor and Picard, 2005; DMello et al., 2008). Large amounts of multimodal data is used to train deep learning algorithms to recognize facial expressions related to emotions and cognitive states in large variety of scenarios.

The definition and detection of metacognitive multimodal indicators requires transforming the raw multi-sensory data collected in a meaningful way so that it allows taking decisions, provide indicators of interlocutor's performance, efficiency and preferences (Greene and Azevedo, 2010). This has been done for interaction logs, the records of sequential actions users performed in an interface. Such actions are interpreted as any communicative action, i.e. having certain communicative functions. A set of dialogue acts has been proposed for screen events by translating the human-human communication mechanisms into human-computer interactions as functions of GUI (van Dam, 2006).

Coherence and interaction analysis is applied to analyse think-aloud interviews and prompting interaction transcripts (Ericsson and Simon, 1984); modern natural language processing techniques are used (Bosch et al., 2021). In multimodal interactions that involve speech, taking notes, nonverbal communication and graphical user interface actions, metacognitive strategies are observable via interaction logs, metacognitive experiences - via recorded and tracked behaviour, and metacognitive knowledge - via speech and typed transcripts. The interaction-based approach to measure metacognition that we propose will enable real-time and non-intrusive assessment of all metacognitive aspects – experiences, knowledge and strategies.

3 Modelling Metacognitive Processes in Dialogue Interaction

Metacognitive regulation refers to adjustments individuals make to their processes to help control their task performance, learning and interaction. Metacognitive processes underlie *awareness*, *monitoring*, *reflection* and *regulation* activities (Brown, 1987). Metacognition has *implicit* and *explicit* forms,[1] and is applied to *own* (sender's) and *others* (addressee's) cognitive processes. In human dialogue, metacognitive processes concern reasoning about interlocutors' intentions and knowledge, and are often modelled as parts of *shared* or *mutual* beliefs forming a *common ground* (Traum, 1994; Bunt, 2000). Common ground is not directly accessible. An access to self and others cognitive processes through questionnaires and think-aloud protocols is very limited; reports on own and others' intentions can be inaccurate. (Meta)cognitive processes underlying establishing and updating common ground (grounding), on the other hand, may

[1] Explicit metacognition is considered a uniquely human ability (Frith, 2012).

become accessible through or inferred from observable dialogue behaviour. For instance, gaze (re-)direction deliver information about the interlocutor attention by means of frequency and duration of gaze fixation on the Areas of Interest (AoI), but also provides an evidence about the positive versus negative emotional reaction on the fixated object. In face-to-face conversation, participants may present evidence of grounding through verbal and vocal signals, body movements and facial expressions; in interaction with graphical user interfaces, typing behaviour, mouse movements and clicks may signal changes in (meta)cognitive and motivational functioning. A metacognitive event is characterised through evidence of reflexive activities indicating any level of sender's mindful awareness about own (sender's) and others (partner's) cognitive process(-es):

- **Level 0:** ignore or offer false continuation;
- **Level 1:** pay and secure attention (mutual eye contact);
- **Level 2:** recognise, record change and respond with minimal signals (gaze (re-)direction, head nods, 'mmhmm', 'uhu'), check out and verify recognition;
- **Level 3:** interpret, check out and verify understanding, and respond to content and feeling ('I see what your mean...', 'I am confused...');
- **Level 4:** evaluate content and feeling, inspect/compare past experiences and verify hypotheses ('I am as worried as you are...');
- **Level 5:** regulate and align, correct/adjust, imitate, anticipate consequences, plan the ongoing procedure (content, sequences, timing,...).

At all these levels, positive and negative beliefs concern sender's awareness about: (i) his/her own thoughts (zero-order theory of mind abilities, Premack and Woodruff (1978)), (ii) about another person's thoughts (first-order theory of mind), and (iii) what another person thinks about sender's thoughts (second-order theory of mind). Consider the following example[2]:

(1) du1. A: The next train is at 11:02.
du2. B: At 11:02.
du3. A: That's correct.
du4. B: Okay thanks

In 1, A in order to continue the dialogue should know that B understands his utterance $du1$ and believes its content p. B's utterance $du2$ can be considered as such evidence where B is verifying its recognition or even on a higher level – its understanding. So after $du2$, A believes that B believes that p, and that B believes that A believes that p. However, A cannot be certain that B indeed believes that p, since in $du2$ he also seems to offer that belief for confirmation. A's response $du3$ gives that confirmation. At this point A does not yet know whether his utterance has reached B and was well understood. B's next contribution $du4$ provides evidence for that; upon understanding $du4$, A has accumulated the following beliefs:

(2) A believes that p
A believes that B believes that p
A believes that B believes that A believes that p
A believes that B believes that A believes that B believes that p
A believes that B believes that A believes that B believes that A believes that p

or represented as *mutual beliefs* equal to:

(3) A believes that it is mutually believed that p

To classify and model implicit and explicit metacognitive events (acts), the framework of the Dynamic Interpretation Theory (DIT, Bunt (1999)) and the ISO 24617-2 dialogue act annotation standard (ISO, 2012) will be used. DIT has emerged from the study of multimodal human-human dialogues uncovering fundamental principles observed in such interactions. DIT and its subset ISO 24617-2 are open multidimensional dialogue act taxonomies[3]. They are proven to provide theoretically grounded and empirically tested inventory of dialogue acts with fine-grained semantic distinctions presenting the semantic framework for the systematic analysis and computational modelling of multimodal dialogue behaviour in many interactive settings.

Special attention will be paid to *feedback* acts which we assume are crucial for successful recognition of metacognitive events: positive and negative feedback about speaker's own (*auto*-feedback) and the partner's processing (*allo*-feedback) at the five processing levels: attention, perception, interpretation, evaluation and execution (Bunt, 2000). Speaker's *repairs, (self-)corrections, partner completions* and *hesitations* (silent and filled pauses) are assumed to strongly correlate with moments of reflection and may reveal speaker's cognitive confidence. *Managing* allocation of *time*, *turn*, *struc-*

[2] Adapted from (Bunt et al., 2007).

[3] DIT, Release 5.2 and ISO 24617-2, Second Edition are available on https://dit.uvt.nl/

Metacognitive Activity	MCQ dimension	Dialogue Act			Indicators (example)
		Dimension	Function	Qualifier	
Awareness	cognitive (self-)conciseness	Auto-/Allo-Feedback	pos. attention pos. perception neg. attention neg. perception	responsiveness (dis)engagement	nonverbal: gaze, head orientation verbal: backchannels nonverbal: gaze aversion GUI: no activity
		Contact Man.	check indication		vocal: throat clearing nonverbal: leaning forward
Monitoring	cognitive confidence	Auto-/Allo-Feedback	pos./neg. interpretation	interest confusion	nonverbal: eye contact nonverbal: puzzled look
		Time Management	stalling	(un)certainty	verbal: filled pauses speech/GUI: slowing down
		Own Communication Management	retraction		verbal/speech: editing expressions GUI: back to initial position speech: disfluencies all: false/re- starts
Reflection	pos./neg. evaluation beliefs	Auto-/Allo-Feedback	pos./neg. evaluation elicitation	empathy worry respect surprise appraisal	nonverbal: thinking face, gaze up verbal: check out understanding verbal: paraphrases, summarization nonverbal: longer gesture strokes verbal: chunking/sorting content nonverbal: raise eyebrow, jerk verbal: make sense, right
Regulation	cognitive need for control	Auto-/Allo-Feedback	pos./neg. execution	irritation cooperation	nonverbal: thinking face, gaze up all: entrainment/alignment
		Own Communication Management	self-correction	frustration excitement	verbal/speech: replacement GUI: cancel
		Partner Communication Management	correct misspeaking completion		verbal: replacement verbal: completion hypothesis
		Discourse Structuring	topic shift		verbal: introduce another topic
		Turn Management	take, keep, release, grab		verbal: start, keep, stop speaking

Table 1: Tentative mapping between metacognitive actions, associated MCQ dimensions and DIT/ISO24617-2 dialogue acts illustrated with examples of possible multimodal metacognitive indicators.[4]

turing discourse and *control* over issues under *discussion* concern with planning aspects. Analysing socio-emotional aspects will enable modelling of metacognitive activities related to positive, negative thoughts, feelings of uncontrollability and danger, and engagement related emotions such as boredom, enjoyment and frustration. Table 1 provides a preliminary view on associations/correlations between metacognitive activities, MCQ dimensions and DIT/ISO dialogue acts illustrated with multimodal behaviour examples. The typology will be experimentally tested and extended as described in the next Section.

4 Experimental Design

Use case The importance of metacognition has been empirically proven for negotiations (Galluccio and Safran, 2015). High self- and others- monitors are more concerned that their negotiations go well, flexibly modify their actions to better adapt to the changing dynamics of the situation, typically by using other people's behaviour as a guide to their own. High self-monitors and -assessors are more likely to engage in argumentation and are better able to accomplish their goals.

As the use case, we will focus on patient-physician negotiations for shared decisions. Medical students and professionals tend to overestimate the value of medical knowledge and are known as poor self-monitors and self-assessors (Eichbaum, 2014). Therapy planning scenarios of varied complexity will be defined reflecting different participant's dispositions. Interaction concerns multi-issue bargaining where each issue involves multiple negotiation *options* with preferences representing parties negotiation positions. Preferences are weighted in order of importance (strength) and defined as the participant's beliefs about *attitudes* towards certain behaviour and *abilities* to perform this behaviour. The goal of each partner is to find out preferences of each other and to search for the best possible mutual agreement. The human participant - doctor - negotiates either with a human or artificial patient who will have different preferences and instructed (programmed) to apply several negotiation and decision-making strategies (Petukhova et al., 2019).[4]

Data Collection will be performed via a) human-human role-playing (small-scaled) and b) human-agent interactive simulations (large collections). Role-playing method is often used to collect

[4]The list of multimodal indicators is not complete, for more examples see (Petukhova, 2005).

interactive data in a controlled setting and underpins simulations of many real-life communicative situations (Brône and Oben, 2015). Here, one participant will be randomly assigned the role of a doctor, the other participant - a patient. Each participant will receive instructions and preference profile, and asked to negotiate a mutual agreement with the highest possible value. Procedures will be specified for the settings where both participants: (1) observe others' actions and flag problems or gaps; (2) verbalise their cognitive processes and their understanding of partner's states and processes; (3) explain his/her choices; and (4) are involved in free flow negotiation. The former three settings will be used as reference for the analysis of metacognition in the unconstrained close to authentic interactions.

Simulations of communicative situations with human and artificial Simulated Patients (SPs) will be arranged. Regular medical communication practice often takes place in a patient-simulated setting, where Simulated Patients (SPs) are involved to portray a particular set of symptoms or roles (Kaplonyi et al., 2017). Simulations with humans provide high fidelity training, but are costly, difficult to reproduce and access. AI agents as SPs can be used to create specific situations in which physicians metacognitive processes can be activated and assessed (Petukhova et al., 2019). Moreover, simulations will impose certain restrictions in order to investigate a controlled set of communicative (metacognitive) activities and related phenomena without having to deal with unrelated details. Multimodal data will be recorded. The quality of recordings will be adapted to the application conditions, i.e. a fairly good but not perfect acoustic and visual quality will be targeted. Prior to recordings, participants will complete the short MCQ-30 questionnaire. We will account for gender and role differences.

Data Recording and Processing Participants' speech will be transcribed by running the Kaldi-based[5] Automatic Speech Recognizer re-trained on the medical in-domain data and correcting the output manually. Since substantial deviations in patient and physician vocabularies are assumed, language models will be adapted to both groups. OpenSMILE tool[6] will be used to extract spectral, prosodic and voice quality features. OpenFace tool[7] will be used to extract 2D/3D facial landmark points for eyes, eyebrows, nose, mouth, jawline and head, and to compute 18 Facial Action Units (AUs). OpenFace enables real-time online/off-line feature extraction from a webcam input and videos, thus no expensive sensors and tracking devices are required. To record GUI interactions, a graphical utility `Atbswp` in Python3 will be used to record the mouse and keyboard actions.

Multi-sensory data will be synchronised and stored in the standard `tei` format, and exported to ELAN[8] to perform the ISO 24617-2 compliant annotations.

5 Expected Outcomes

The proposed project will contribute to a better understanding of metacognitive processes underlying dialogue participants decision-making and interactive performance progressing towards a computational cognitive model of social metacognition. An interaction-based method for metacognition assessment will be worked out providing an ISO-compliant typology of metacognitive events, a set of multimodal feature extraction and classification models as well as new tools for multidimensional dialogue analysis. Finally, substantial amount of multimodal data annotated with the ISO 24617-2 dialogue acts will be provided to the research community via DialogBank release. [9]

References

Maria Bannert and Christoph Mengelkamp. 2008. Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. does the verbalisation method affect learning? *Metacognition and Learning*, 3(1):39–58.

Nigel Bosch, Yingbin Zhangm, Luc Paquette, Ryan Baker, Jaclyn Ocumpaugh, and Gautam Biswas. 2021. Students verbalized metacognition during computerized learning. In *ACM SIGCHI: Computer-Human Interaction*.

Geert Brône and Bert Oben. 2015. Insight interaction: a multimodal and multifocal dialogue corpus. *Language resources and evaluation*, 49(1):195–214.

Ann Brown. 1987. Metacognition, executive control, self-regulation, and other more mysterious mechanisms. *Metacognition, motivation, and understanding*.

Harry Bunt. 1999. Dynamic interpretation and dialogue theory. *The structure of multimodal dialogue*, 2:139–166.

[5] https://kaldi-asr.org/

[6] https://www.audeering.com/opensmile/

[7] https://github.com/TadasBaltrusaitis/OpenFace

[8] https://archive.mpi.nl/tla/elan

[9] https://dialogbank.uvt.nl/

Harry Bunt. 2000. Dialogue pragmatics and context specification. *Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics. Amsterdam: Benjamins*, pages 81–150.

Harry Bunt, Roser Morante, and Simon Keizer. 2007. An empirically based computational model of grounding in dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 283–290.

Sam Cartwright-Hatton and Adrian Wells. 1997. Beliefs about worry and intrusions: The metacognitions questionnaire and its correlates. *Journal of anxiety disorders*, 11(3):279–296.

Hans van Dam. 2006. *Dialogue acts in GUIs*. Ph.D. thesis, Technische Universiteit Eindhoven, Department of Industrial Design.

Sidney DMello, Tanner Jackson, Scotty Craig, Brent Morgan, P Chipman, Holly White, Natalie Person, Barry Kort, R El Kaliouby, Rosalind Picard, et al. 2008. Autotutor detects and responds to learners affective and cognitive states. In *Workshop on emotional and cognitive issues at the international conference on intelligent tutoring systems*, pages 306–308.

Quentin G Eichbaum. 2014. Thinking about thinking and emotion: the metacognitive approach to the medical humanities that integrates the humanities with the basic and clinical sciences. *The Permanente Journal*, 18(4):64.

K Anders Ericsson and Herbert A Simon. 1984. *Protocol analysis: Verbal reports as data.* the MIT Press.

Chris D Frith. 2012. The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2213–2223.

Mauro Galluccio and Jeremy D Safran. 2015. Mindfulness-based training for negotiators: Fostering resilience in the face of stress. In *Handbook of International Negotiation*, pages 209–226. Springer.

Dragan Gašević, Shane Dawson, and George Siemens. 2015. Lets not forget: Learning analytics are about learning. *TechTrends*, 59(1):64–71.

Jeffrey A Greene and Roger Azevedo. 2010. The measurement of learners self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational psychologist*, 45(4):203–209.

ISO. 2012. *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2*. ISO Central Secretariat, Geneva.

Jessica Kaplonyi, Kelly-Ann Bowles, Debra Nestel, Debra Kiegaldie, Stephen Maloney, Terry Haines, and Cylie Williams. 2017. Understanding the impact of simulated patients on health care learners communication skills: a systematic review. *Medical education*, 51(12):1209–1219.

Ashish Kapoor and Rosalind W Picard. 2005. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682.

Karolina M Lempert, Yu Lin Chen, and Stephen M Fleming. 2015. Relating pupil dilation and metacognitive confidence during auditory decision-making. *PLoS One*, 10(5):e0126588.

Stephanie B Linek, Birgit Marte, and Dietrich Albert. 2008. The differential use and effective combination of questionnaires and logfiles. In *Computer-based Knowledge & Skill Assessment and Feedback in Learning settings (CAF), Proceedings of the International Conference on Interactive Computer Aided Learning (ICL), 24th to 26th September*.

Andrei Malchanau, Volha Petukhova, and Harry Bunt. 2018. Towards integration of cognitive models in dialogue management: designing the virtual negotiation coach application. *Dialogue & Discourse*, 9(2):35–79.

Judith Meessen, Stefan Sütterlin, Siegfried Gauggel, and Thomas Forkmann. 2018. Learning by heartthe relationship between resting vagal tone and metacognitive judgments: a pilot study. *Cognitive processing*, 19(4):557–561.

Volha Petukhova. 2005. *Multidimensional interaction of multimodal dialogue acts in meetings*. Ph.D. thesis, MA thesis, Tilburg University.

Volha Petukhova, Firuza Sharifullaeva, and Dietrich Klakow. 2019. Modelling shared decision making in medical negotiations: Interactive training with cognitive agents. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 251–270. Springer.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain sciences*, 1(04):515–526.

Gregory Schraw. 2009. A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and learning*, 4(1):33–45.

Jonathan Michael Spector. 2014. Conceptualizing the emerging field of smart learning environments. *Smart learning environments*, 1(1):1–10.

Niels A Taatgen. 2013. The nature and transfer of cognitive skills. *Psychological review*, 120(3):439.

David R Traum. 1994. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science.

Martijn E Wokke, Dalila Achoui, and Axel Cleeremans. 2020. Action information contributes to metacognitive decision-making. *Scientific reports*, 10(1):1–15.

Annotating Quantified Phenomena in Complex Sentence Structures Using the Example of Generalising Statements in Literary Texts

Tillmann Dönicke
Göttingen Centre for Digital Humanities
University of Göttingen
tillmann.doenicke@uni-goettingen.de

Luisa Gödeke
Department of German Philology
University of Göttingen
luisa.goedeke@uni-goettingen.de

Hanna Varachkina
Department of German Philology
University of Göttingen
hanna.varachkina@stud.uni-goettingen.de

Abstract

We present a tagset for the annotation of quantification which we currently use to annotate certain quantified statements in fictional works of literature. Literary texts feature a rich variety in expressing quantification, including a broad range of lexemes to express quantifiers and complex sentence structures to express the restrictor and the nuclear scope of a quantification. Our tagset consists of seven tags and covers all types of quantification that occur in natural language, including vague quantification and generic quantification. In the second part of the paper, we introduce our German corpus with annotations of generalising statements, which form a proper subset of quantified statements.

1 Introduction

Quantification is a core element of human language because it allows us to make statements about groups or classes of entities, in contrast to statements about individually referenced entities.

One subtype of quantified statements are *generalising* or *generic* (for now summarised as *generalising*) statements that involve quantification over assumed members of a class rather than contextually given entities. These generalising statements are particularly interesting for NLP applications that operate on discourse-level, e.g. in knowledge extraction (e.g. Bhakthavatsalam et al., 2020) and argumentation mining (e.g. Becker et al., 2016). But also in computational literary studies, generalising statements can be viewed as indicators (i.e. features) for meta-level phenomena such as passages of (self-)reflection (cf. Lahn and Meister, 2016, p. 184) or passages addressing real-world issues within a fictional text (e.g. Vesper, 2014).

Put very concisely, one traditionally differentiates between (determiner or generic) quantification on NP-level and (adverbial or generic) quantification on clause-level (Krifka and Gerstner, 1987; Partee, 1990; Krifka, 2016), which is also reflected in certain annotation schemes (cf. Friedrich et al., 2015). However, for such higher-level applications, where the presence of quantification or, more specifically, generalisation serves as a feature, the syntactic or semantic structure of quantified statements plays only a subordinate role. Therefore, performing a syntactic and/or semantic analysis during the annotation would be laborious but not expedient—especially in domains where sentences tend to employ a complex syntactic structure, such as literary texts. Moreover, quantified statements are not always marked by an overt linguistic marker but can also be covertly quantified in the case of generic statements.

For generalisation specifically, previous work commonly differentiates between generalisation over (kinds of) individuals and generalisation over recurring events/situations (Krifka et al., 1995; Carlson, 2011; Friedrich and Palmer, 2014). Similar to the syntactic categorisation, this is not only an insufficient differentiation if one is interested in generalising statements as a whole; it further constitutes a limitation, since it is possible to quantify (and thus generalise) over other types of entities than the two just mentioned. We shall expand on the theoretical syntactic and semantic considerations in Section 2 and give an overview of related practical challenges in literary texts in Section 3.

Considering both the theoretical and practical aspects, we developed a tagset and annotation guidelines for generalising statements that are neither bound to syntactic nor to semantic properties and preserve only the information which is most important in our view: the type of quantification (universal/existential/vague etc.). This shallow annotation scheme allows a comparatively fast annotation of generalising statements, which is especially valu-

Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pages 20–32

Quantifier	Restrictor	Nuclear scope
determiner, adverb, negation, generic	subordinate clause, if-clause, common NP, topic	main clause, assertion, predication, focus

Table 1: Examples for tripartite-structure components (Partee, 1990, p. 10)

able in the literary domain. Since our annotation scheme can be used for quantified statements in general, we will present it as such in Section 4 and turn to generalising statements in Section 5. Afterwards, Section 6 presents details about our corpus and annotation process. Sections 7 and 8 discuss related and future work, respectively.

2 Quantification

In English (and in German), quantificational notions are typically triggered by determiners, e.g. *all*, *most* etc., or adverbs, e.g. *always*, *usually* etc. Following Lewis (1975), Kamp (1981) and Heim (1982), we assume quantified statements to consist of a tripartite logical form consisting of the quantifier Q, the restrictor and the nuclear scope:

(1) $Q[x : restr(x)][scope(x)]$

We subsume determiners and adverbs to both function as quantifiers in this model:

(2) a. Most horses have four legs.
b. MOST$[x : \text{horse}(x)][\text{four-legged}(x)]$

(3) a. Usually, a horse has four legs.
b. USUALLY$[x : \text{horse}(x)][\text{four-legged}(x)]$

This approach enables us to include various clause-level forms of quantification into a unifying analysis. The forms can differ in syntactic realisation; Table 1 shows some examples for how quantifier, restrictor and nuclear scope can be realised in natural language, beyond determiners/adverbs and common noun phrases.

In addition to the syntactic diversity of quantification, it can range over all types of semantic entities. While quantification over individuals, as in (2) and (3), and events/situations, so-called *habituals* (e.g. Rimell, 2004) as in (4), may be most notable, it is also possible to quantify over e.g. times, as in (5), and locations, as in (6).

(4) a. John usually drives to work
b. USUALLY$[e : \text{agent}(e, j) \wedge \text{to.work}(e)][\text{drive}(e)]$

(5) a. It snows every winter
b. $\forall[t : \text{winter}(t)][\text{GEN}[e : e \subseteq t][\text{snow}(e)]]$

(6) a. It snows in Austria
b. GEN$[l : \text{in.Austria}(l)][\text{GEN}[e : \text{location}(e, l)][\text{snow}(e)]]$

3 Challenges in Literary Texts

We are interested in the distribution of such structures as "general statements" or "statements of universal validity" in German fictional texts. Therefore, we are aiming at annotating quantified statements in any form they may occur in. Our corpus consists of fictional texts written or published between 1650 and 1950. Hence, we are not only confronted with complex sentence structures, which are typical for literary texts, but also with older versions of German. We need an annotation concept that lets us capture quantified expressions in all their variety. Although our research does not primarily focus on the surface quantification, but on the generalising function that these structures fulfil, our work has a solid foundation in (formal) linguistics, as we will show in Section 4. Therefore, the transfer of theoretical knowledge about quantification into computational linguistics turns out to be a challenge—especially for analysing the literary domain. Particularly, we are facing three main challenges: First, the default formal analysis of quantification by defining quantifier, restrictor, and scope (as established in the previous section) can be highly complex in sentences of fictional writing. On top of that, German allows comparatively long and complex multi-clause sentences, especially in older language variants. This issue is illustrated in (7), where we already cut out several embedded if-clauses (German: *wenn* 'if') out of this one sentence. The English translation in (7') does not take the various if-clauses on and splits them up into separate sentences:

(7) Wenn Luciane, meine Tochter, die für die Welt geboren ist, sich dort für die Welt bildet, [...]; wenn sie durch Freiheit des Betragens, Anmut im Tanze, schickliche Bequemlichkeit des Gesprächs sich vor allen auszeichnet und durch ein angebornes herrschendes Wesen sich zur Königin des kleinen Kreises macht, wenn

die Vorsteherin dieser Anstalt sie als kleine Gottheit ansieht, die nun erst unter ihren Händen recht gedeiht, die ihr Ehre machen, Zutrauen erwerben und einen Zufluß von andern jungen Personen verschaffen wird, wenn [...]: so ist dagegen, was sie schließlich von Ottilien erwähnt, nur immer Entschuldigung auf Entschuldigung [...]. (Goethe, WV)

(7') Luciana, my daughter, born as she is for the world, is there training hourly for the world [...] She distinguishes herself above every one at the school with the freedom of her carriage, the grace of her movement, and the elegance of her address, and with the inborn royalty of nature makes herself the queen of the little circle there. The superior of the establishment regards her as a little divinity, who, under her hands, is shaping into excellence, and who will do her honor, gain her reputation, and bring her a large increase of pupils; [...] while her concluding sentences about Ottilie are nothing but excuse after excuse. (Goethe, EA, p. 23 f.)

If-clauses, as in (7) can be considered as restrictors (compare Table 1); and the then-clause (German: *so* 'then') can be considered as nuclear scope. Our first problem manifests itself here: The if-clauses form a list of coordinated restrictors for only one scope, and it remains unclear how many individual quantified statements there are or whether the individual restrictors are meant to be joined by logical conjunction or disjunction. Resolving such issues would require a laborious and—in our case—redundant analysis.

For ease of presentation, we shall only use English examples in the following, taken from official translations. The original examples are provided in the appendix (B).

Second, we have to deal with ambivalent syntactic structures, leading to scope ambiguity. If a sentence carries more than one quantifier, different readings arise due to the dominant quantification.

(8) Help upon the spot is the thing you often most want in the country. (Goethe, EA, p. 49)

In (8) we find two generic expressions (*help upon the spot* and *the country*) combined with the adverbial *often*. Third, the absence of overt markers for quantification is a greater problem than an overpresence. The generic NPs (cf. Leslie and Lerner, 2016), e.g. *business* and *life* in (9), certainly have a generalising function in this context, but are not overtly quantified.

(9) Business requires earnestness and method; life must have a freer handling. (Goethe, EA, p. 46)

In the following section, we will present our annotation tagset, which allows us to tackle these issues.

Tag	Description
ALL	overt universal quantification
MEIST	overt majority quantification
EXIST	overt existential quantification
ZAHL	overt numerical quantification
DIV	overt vague quantification
BARE	none of the above + covert quantification
NEG	any of the above + negation

Table 2: Tagset

4 A Tagset for Quantified Statements

The complete tagset is summarised in Table 2. Because of the challenges associated with identifying restrictor and scope of a quantified statement, we do not annotate them separately. Instead, we label the whole span which contains quantifier, restrictor and scope. The tags in our tagset represent the (semantic) type of quantification. We take clauses as the smallest unit of annotation, meaning that one quantified statement may comprise one clause, as in (10), or several clauses, as in (11–12). Punctuation at annotation boundaries is omitted.

(10) [Most horses have four legs]$_{\text{MOST}}$.

(11) [A whale which is ill yields no blubber]$_{\text{BARE}}$. (cf. Burton-Roberts, 1976)

(12) [He who gets up early gets tired quicker]$_{\text{BARE}}$.

We use brackets to indicate annotation spans and subscripts to denote tags. The following subsections motivate the individual tags.

4.1 Precise Quantification

Natural language employs a clear-cut set of mathematically precise quantifiers, whose meanings can be defined using set relations (see Table 3). All

Name	Q	$Q[x : restr(x)][scope(x)]$ **iff**
universal	$\forall$	$\lvert S_{restr} \cap S_{scope}\rvert = \lvert S_{restr}\rvert$
majority	MOST	$\lvert S_{restr} \cap S_{scope}\rvert > \lvert S_{restr} \setminus S_{scope}\rvert$
existential	$\exists$	$\lvert S_{restr} \cap S_{scope}\rvert \neq 0$
counting	$\exists^{Rn}$	$\lvert S_{restr} \cap S_{scope}\rvert\ R\ n$
proportional	$Q^{Rn/m}_{prop}$	$\lvert S_{restr} \cap S_{scope}\rvert\ R\ n/m \cdot \lvert S_{restr}\rvert$

Table 3: Truth conditions for precise quantifiers; $S_P := \{x : P(x)\}$ is the extension of P

of these quantifiers are expressed by a number of lexemes at the surface of a sentence. $\forall$ is expressed by *all*, *every*, *always*, *everywhere* etc., and MOST usually appears as *most(ly)* or *main(ly)*. Statements with these quantifiers should be labelled with the tags ALL and MEIST (German for "MOST"), respectively:

(13) There is lime, you remember, **[**which shows the strongest inclination for all sorts of acids—a distinct desire of combining with them**]**$_{\text{ALL}}$. (Goethe, EA, p. 55 f.)

(14) **[**Men think most of the immediate—the present**]**$_{\text{MEIST}}$; (Goethe, EA, p. 16)

$\exists$ is associated with the indefinite article *a/an* in classical Fregean semantics (Zalta, 2020), as (15a) and (15b) exemplify. Fodor and Sag (1982) note, however, that a statement as in (15a) rather is ambiguous between the quantified interpretation in (15c) and the referential interpretation in (15d) (cf. von Heusinger, 2000)[1]. We follow this analysis and do not consider indefinite NPs to be markers for existential quantification. Instead, statements with a meaning as in (15c) are treated as genuine generic quantification (see Section 4.4): and statements with a meaning as in (15d) must not be labelled since they do not contain quantification.

(15) a. A man walks
b. $\exists[x : \text{man}(x)][\text{walk}(x)]$
c. $\text{GEN}[x : \text{man}(x)][\text{walk}(x)]$
d. $\text{walk}(\varepsilon_i x\, \text{man}(x))$

We use the tag EXIST for explicit existential statements instead. In English, such statements can be formulated with the expression *there is/are* or the verb *exist*:

[1]The expression $\varepsilon_i x\, P(x)$ returns an entity which satisfies the predicate P, based on the choice function i (Avigad and Zach, 2020).

(16) **[**Thirdly, there are those people who investigate the sea bed as if it were a meadow**]**$_{\text{EXIST}}$. (Fontane, Stechlin, p. 288)

(17) **[**But they still do exist, they've got to exist or else they've got to exist *again***]**$_{\text{EXIST}}$. (Fontane, Stechlin, p. 130)

There are different theories on how to analyse such existential statements, differing in the question whether existence is a quantifier or a predicate and, in case of the latter, what kind of predicate it is (McNally, 1998; Moltmann, 2013). Although we do not have a preference for either analysis, we can observe that the verb *exist* must sometimes be analysed as a scope predicate rather than a quantifier. For example, it could be analysed either as existential quantifier or predicate of a covert quantifier in (18), whereas the quantifier analysis is not possible in (19) because of the overt quantifier MOST. The difference between (18b) and (18c) is very subtle and it is not always easy or even possible to identify the correct analysis—especially because a generic quantifier can also have an existential interpretation (Cohen, 2004). Therefore, and because we prefer to keep all occurrences of *exist* in one class, we label all occurrences with EXIST and treat cases like (19) as double quantification (see Section 4.3).

(18) a. **[**Fairy-tale creatures exist**]**$_{\text{EXIST}}$
b. $\exists[x][\text{fairy-tale.creature}(x)]$
c. $\text{GEN}[x : \text{fairy-tale.creature}(x)][\text{exist}(x)]$

(19) a. **[**Most fairy-tale creatures exist**]**$_{\text{EXIST+MEIST}}$
b. $\text{MOST}[x : \text{fairy-tale.creature}(x)][\text{exist}(x)]$

The last type of precise quantifiers are numerical quantifiers, which either express absolute counts ($\exists^{Rn}$) or proportions ($Q^{Rn/m}_{prop}$). Numerical quantifiers are composed of numerals, such as *one*, *two*, *half*, *third*, *dozen*, *hundred*, *percent*, *million*, corresponding to n (and m). Numerals are optionally combined with an expression like *at least*, *exactly*, *up to* etc., corresponding to a mathematical relation $R \in \{=, <, >, \leq, \geq, ...\}$:

(20) a. Five men walk
b. $\exists^{=5}[x : \text{man}(x)][\text{walk}(x)]$

(21) a. At least five men walk
b. $\exists^{\geq 5}[x : \text{man}(x)][\text{walk}(x)]$

(22) a. Up to two thirds of men walk
b. $Q^{\leq 2/3}_{prop}[x : \text{man}(x)][\text{walk}(x)]$

Numerical quantification should be labelled with the tag ZAHL (German for "NUMBER"):[2]

(25) The county had always gone Conservative and it was a matter of honor to go Conservative again, as Luther had said, "[Even if the world were full of a thousand devils]$_{\text{ZAHL}}$." (Fontane, Stechlin, p. 140)

The reader might argue that "almost all" as in (26) also has a mathematical definition (that is "all but finitely many") and should thus receive a separate tag. In natural language, however, *almost* modifies the truth value of a statement rather than its quantification (Kilbourn-Ceron, 2014). In fact, *almost* can appear in combinations with other quantifiers (see (27)) and without any quantifier (see (28)) as well, hence we do not include a separate tag for "almost all" in our tagset. We also treat similar modifiers like *hardly*, *nearly*, *more or less* etc. as not affecting the type of quantification.

(26) [And so for starters he's got to conquer everything, almost all the towns roundabout and all the castles for sure]$_{\text{ALL}}$. (Fontane, Stechlin, p. 82)

(27) [Almost five men walk]$_{\text{ZAHL}}$

(28) "You must let me make what will seem a wide sweep; we shall be on our subject almost immediately." (Goethe, EA, p. 53)

4.2 Vague Quantification

In addition to the precise quantifiers, one can find a broad range of vague quantifiers in natural language, whose truth conditions cannot be defined precisely. Some lexemes are *few*, *some*[3], *many*, *rarely*, *occasionally*, *commonly*, *often*; and multiword expressions like *as a rule* or *in general* can also express vague quantification. The vagueness makes it difficult to determine how many semantically different quantifiers there are—if not every lexeme represents its own quantifier. For example, is *often* the same as *frequently*? We therefore group all vaguely quantified statements under the tag DIV (for "diverse"):

(29) "[Our excellent superior commonly permits me to read the letters in which she communicates her observations upon her pupils to their parents and friends]$_{\text{DIV}}$. (Goethe, EA, p. 43)

(30) "It concerns our friend the Captain," answered Edward; "you know the unfortunate position [in which he, like many others, is placed]$_{\text{DIV}}$. (Goethe, EA, p. 13)

4.3 Multiple Quantification

As mentioned in Section 3, several quantifiers can occur within a statement; or several quantified statements can be nested as in (31). Annotations from overlapping statements do not affect each other, hence this is no multi-label case in the sense of having multiple tags for one statement. It can become a multi-label case if one merges tags on token or clause level, though, e.g. for measuring annotator agreement or evaluating a quantification tagger.

Statements that contain more than one overt quantifier, on the other hand, should receive all corresponding tags, as in (32). We treat the assigned tags as a set, meaning that every of the five tags for overt quantification can be assigned only once to a statement, even if there are e.g. several quantifiers qualifying for the ALL tag, as in (33).

(31) His entrance into the regiment more or less coincided with the beginning of the reign of Friedrich Wilhelm IV, [and whenever he mentioned that fact, he took pleasure in poking a bit of fun at himself by stressing that "[all great events have their accompanying secondary phenomena]$_{\text{ALL}}$."]$_{\text{ALL}}$

[2] A potentially conflicting case is $Q^{>1/2}_{prop}$, which is mathematically equivalent to MOST. Following Hackl (2009), who provides evidence for a cognitive difference between *more than half* and *most*, we label these expressions as follows:

(23) a. [Most of the men walk]$_{\text{MEIST}}$
b. $\text{MOST}[x : \text{man}(x)][\text{walk}(x)]$

(24) a. [More than half of the men walk]$_{\text{ZAHL}}$
b. $Q^{>1/2}_{prop}[x : \text{man}(x)][\text{walk}(x)]$

[3] The authors of this paper intensely discussed whether the German *manch*, which has a similar meaning to that of *some*, should be DIV or EXIST, because some scholars analyse *manch*/*some* as existential quantifier (e.g. Löbner, 2005; Chierchia and McConnell-Ginet, 2000, p. 310). We decided to label *manch* with DIV since it usually implies an indefinite but substantial (vague) number of entities (Dudenredaktion, n.d.). (Furthermore, the determiner does not quite fit in the EXIST class from a morphological perspective.) We suggest that one should classify statements with *some* as DIV by the same argument. Note that *manch* always causes a quantificational interpretation (with singular and with plural NPs), whereas *some* can also cause a referential interpretation (like the indefinite article in (15d)) when combined with a singular NP (Winter, 1997). In the latter case, no tag would be assigned.

(Fontane, Stechlin, p. 3)

(32) **[**Whoever eats meat sometimes is a murderer forever**]**$_{\text{ALL+DIV}}$

(33) **[**Every Pope loves all his subjects equally**]**$_{\text{ALL}}$

The individual quantifiers within a statement do not always employ an unambiguous hierarchy, or the hierarchy becomes apparent after labourious semantic analysis only. Therefore, we do not opt for an ordering of the tags in the case of multiple quantification.

There are more morphosyntactically complex quantifiers in natural language than we could discuss in the previous subsections (see Keenan and Paperno (2012) for an extensive overview). Complex quantifiers should be decomposed whenever no single tag is applicable, which can also result in a multi-label annotation:

(34) **[**All but two men walk**]**$_{\text{ALL+ZAHL}}$

4.4 Generic Quantification

In opposition to the other quantifiers discussed so far, the generic quantifier GEN is covert, i.e. it is not marked by a specific lexical item. Instead, there is a broad range of surface forms that can mark genericity. The statements in (35), for example, (cf. Carlson, 2011) all make a general claim about lions.

(35) a. The lion is ferocious
b. Lions are ferocious
c. A lion is ferocious
d. $\text{GEN}[x : \text{lion}(x)][\text{ferocious}(x)]$

While (35) shows generic statements about entities, (36) shows generic statements about events. Note that we only show one possible analysis in (36c), although several interpretations are possible due to scope ambiguities.

(36) a. John eats meat
b. John used to eat meat[4]
c. $\text{GEN}[e : \text{eat}(e) \wedge \text{agent}(e, j)][$
$\text{GEN}[y : \text{meat}(y)][\text{patient}(e, y)]]$

Consecutively, (37) is a generic statement over both entities and events:

(37) a. **[**Lions eat meat**]**$_{\text{BARE}}$
b. $\text{GEN}[x : \text{lion}(x)][$
$\text{GEN}[e : \text{eat}(e) \wedge \text{agent}(e, x)][$
$\text{GEN}[y : \text{meat}(y)][\text{patient}(e, y)]]]$

We are aware that some semanticists would replace some of the generic quantifications in (37b) by existential quantifications or even non-quantificational expressions. This illustrates, however, how difficult it is to find covert generic quantifiers compared to overt quantifiers as in (38).

(38) a. **[**Most lions always eat some meat**]**$_{\text{ALL+DIV+MEIST}}$
b. $\text{MOST}[x : \text{lion}(x)][$
$\forall[e : \text{eat}(e) \wedge \text{agent}(e, x)][$
$\text{SOME}[y : \text{meat}(y)][\text{patient}(e, y)]]]$

With increasing complexity of sentence structures—as in fictional texts—, it is simply impossible to determine all covert quantifiers unambiguously in the annotation process. However, if no overt quantifier appears in a statement and the statement still has a quantificational meaning then there must be a covert quantifier somewhere. Hence quantified statements without any overt quantifier should be labelled with the tag BARE:[5]

(39) **[**The country people have knowledge enough**]**$_{\text{BARE}}$, **[**but their way of imparting it is confused**]**$_{\text{BARE}}$, **[**and not always honest**]**$_{\text{NEG}}$. **[**The students from the towns and universities are sufficiently clever and orderly, but they are deficient in personal experience**]**$_{\text{BARE}}$.

4.5 Negation

Negation can occur in different syntactic positions and cause problematic cases during the annotation. If the quantifier or the scope in a universally or existentially quantified statement is negated, its meaning could be expressed as both a universal or existential quantification, following the negation rules for quantifiers:

(40) $\neg\forall[x : restr(x)][scope(x)]$
$\equiv \exists[x : restr(x)][\neg scope(x)]$

[4] Ignoring tense. The German *pflegen zu* 'use to' can also be used in present tense; unfortunately, there seems to be no equivalent present-tense construction in English.

[5] It might be confusing why the passage about the country people in (39) is fragmented whereas the passage about the students is not. According to our annotation guidelines, two or more subsequent quantified statements should be joined if they receive the same tag and the restrictor or the scope stay the same. This condition is not fulfilled for the former passage where the restrictor firstly shifts from *country people* to *their way of impairing it* and the tag secondly changes from BARE to NEG.

(41) $\neg\exists[x : restr(x)][scope(x)]$
$\equiv \forall[x : restr(x)][\neg scope(x)]$

The case becomes even more complicated with ambiguous negation lexemes. The determiner *no* could be analysed as $\neg\exists$ or $\neg$ GEN, hence the statement in (42a) could be analysed as both (42b) and (42c).

(42) a. **[**No lion sleeps**]**$_{\text{NEG}}$

b. $\neg\exists[x : \text{lion}(x)][\text{sleep}(x)]$
$\equiv \forall[x : \text{lion}(x)][\neg\text{sleep}(x)]$

c. $\neg\,\text{GEN}[x : \text{lion}(x)][\text{sleep}(x)]$
$\overset{?}{\equiv}\ \text{GEN}[x : \text{lion}(x)][\neg\text{sleep}(x)]$

This means that one could find arguments to label (42a) with any of EXIST, ALL or BARE. Resolving (ambiguous) negation would require to know whether it applies to the quantifier, restrictor or scope of a statement, and a set of detailed definitions for how one should annotate cases as in (40–42). Again, such a detailed analysis does not fit our aim of developing a simple annotation procedure. The simplest solution to this issue is to assign a special tag NEG for quantified statements with any negation in it. NEG then replaces all other tags that one could assign:

(43) "**[**And there are many cases [...] in which we are obliged, and in which it is the real kindness, rather to write nothing than not to write**]**$_{\text{NEG}}$." (Goethe, EA, p. 20)

5 Generalising Interpretations

In the previous section, we presented a tagset for quantification. However, our research does not focus on quantified statements in general but only on generalising statements, which we consider to be a subset of quantified statements. The main purpose of our research is to find generalising statements in fictional works of literature to investigate their narratological function. Our working definition for generalisation results from previous work on the re-interpretation of universal quantifiers: Löbner (2005) already notes two interpretations for *every* (originally the German counterpart *jede*) as in (44), namely 1) a concrete quantification over contextually determined instances, and 2) a generic quantification over assumed instances. Similarly, Leslie et al. (2011) found that adults frequently judge universal statements as in (45) true, despite knowing that there are counterexamples. Leslie et al. (2011) conclude that *all* can be interpreted as a generic quantifier instead of a universal quantifier, and calls this the "generic overgeneralisation effect".

		Trueness erroneously accepted	
		no	yes
Instances assumed	no	(i) All students in the semantics class take notes	(ii) All students in the semantics class are broke
	yes	(iii) All triangles have three sides	(iv) All ducks lay eggs

Table 4: Quantified statements with varying characteristics

(44) Every child is entitled to a place in school

(45) All ducks lay eggs

According to these works, an "overgeneralised" (universal) statement seems to be characterised by two properties:

1. The quantification involves assumed instances, i.e. not all restricted instances are contextually determined.
2. The statement is accepted as true (in the context of utterance) although there is not enough evidence for its trueness (because of unknown instances), or there is evidence for its falseness (because of known counterexamples).

Quantified statements that fulfil both properties are clearly generalising whereas statements fulfilling none of them clearly are not. Statements that fulfil only one of the properties are harder to classify, which is why we want to briefly discuss them in the following.

For examples (i) and (ii) in Table 4, imagine a classroom situation in which all students take notes. Then there is no doubt that (i) is true. Furthermore, some (but not all) of the students look the worse for wear; hence there is not enough evidence to claim (ii) as being true. Thus, one could call (ii) a "generalisation" from a subset of the students to all of them, taking the ordinary meaning of "generalisation" into account. However, the additional (cognitive) process of generalising to any assumed instances is missing, which is why it is no generalising statement in our sense.

The restrictor in (iii) and (iv) includes all triangles or ducks, respectively. Hence the quantified instances include assumed/unknown instances. Still, even for triangles which one does not know, one can infer that they have three sides (because otherwise they would not be triangles by definition). Therefore, we do not consider such *(quasi-) definitional* statements (Leslie et al., 2011) to be generalising.[6]

Having only looked at universal quantification so far, we now claim that generalisation can occur with every natural-language quantifier, and frequently does so in literary texts. Many of the examples shown in previous sections are in fact generalising; additional examples for each tag are given in the appendix (A). Unlike Leslie et al. (2011) and Löbner (2005), we do not claim that the quantifier in a generalising statement is re-interpreted as generic quantifier. Based on our observations (cf. Sec. 6), we suppose that generic quantification is a frequent but not the only type of quantification used to express generalisation.

6 Corpus and Annotations

We currently construct a diachronic corpus of German fictional literature from 1650 to 1950. As of now, we annotated generalising interpretations in ten texts—excluding some additional texts from a pilot annotation for developing our annotation guidelines. CATMA[7] appeared to be most suitable for annotating fictional texts and became our tool of choice. In order to create a versatile dataset and save resources, we annotate only the beginning of every text (usually the first about 200 sentences).

Our annotation procedure is as follows: Each text is first annotated by two out of six student assistants. In a second step, two researchers glance over the text again, focusing on the statements that were annotated by at least one annotator, discuss the annotations and create an expert annotation as gold standard. Arguably, this procedure is prone to false negatives, i.e. statements that none of the annotators identified are likely to be missed while creating the gold standard.

We measure inter-annotator agreement on token-level (excluding punctuation) using κ (Fleiss et al., 2003), treating the occurring tag combinations as classes (i.e. none, ALL, ..., ALL+DIV, ...). Since the annotators vary between texts, we first compute the agreement separately for each text. The average agreement is shown in Table 5. On average, there is substantial agreement of 0.67 for annotating generalising statements (GI+Q). The relatively high standard deviation of 0.20 indicates that there is a great variance between texts and/or annotators. Additionally, we compute the agreement when just distinguishing between "no tag" and "any tag", i.e. "not generalising" vs. "generalising" (GI). The values are almost identical to those for GI+Q, indicating that the overall agreement is mainly influenced by the agreement on what is a generalising interpretation. To estimate the applicability of our tagset, we also compute the agreement for only those tokens that received a tag by all annotators, i.e. those tokens where the annotators agreed that they are part of a generalising statement (Q). Here, we see an average agreement of 0.85, which is significantly higher than that for GI, indicating that annotating quantification is comparatively straightforward.

GI+Q		**GI**		**Q**	
κ	σ	κ	σ	κ	σ
.67	.20	.68	.22	.85	.14

Table 5: Mean inter-annotator agreement (κ) over all texts and corresponding standard deviations (σ); see text for column meanings

Generalising statements						
ALL	MEIST	EXIST	DIV	BARE	NEG	Total
151	7	17	76	332	145	728

Table 6: Number of generalising statements in all texts

Annotation results can be found in Table 6. Within 2,791 annotated sentences (61,979 tokens), 728 generalising statements occur (in the gold standard), which have an average length of 17 tokens. We can see that most generalising statements use plain generic quantification (BARE); followed by universal quantification (ALL) and vague quantification (DIV). Generalising statements with existential quantification (EXIST) or majority quantification (MEIST) are far less common. Note, however, that the counts do not directly reflect a "generalisation potential" of the individual quantification types since some quantification types occur with a higher frequency than others in the first place. We removed the tag ZAHL from our annotation

[6] *Generic* is not the same as *generalising* by our definition: Since generic quantification can also be used to express definitional statements (e.g. *triangles have three sides*), not all generic statements are generalising statements.

[7] https://catma.de/

guidelines because generalising interpretations for numerical quantifiers hardly occurred during the pilot annotation (and ever since then).

7 Related Work

While the data in the existing corpora that are annotated with phenomena related to generalisation usually originates from the domains traditional for NLP, such as news and internet communication, we investigate generalisation in fiction, a more complex domain. To the best of our knowledge, there are no corpora on generalisation or genericity for German and our work reduces this gap.

Most existing resources focus on noun phrases, often in the context of coreference resolution. A detailed survey on generics in the coreference resolution research can be found in Nedoluzhko (2013).

Friedrich et al. (2015) provide a survey of genericity-annotated corpora for English. They note that ACE corpora (Mitchell et al., 2003; Walker et al., 2006) are most widely used, e.g. by Reiter and Frank (2010), to identify generic noun phrases using a supervised approach.

Friedrich et al. (2015) were the first to suggest an annotation scheme for generic statements where both clauses and their subject NPs are annotated with the labels "generic" and "non-generic". However, they only consider kind-referring generics and exclude e.g. habitual statements. In a subsequent work, Friedrich et al. (2016) investigate both habituals and kind-referring generics as two separate situation entity types, alongside with states and events, in a sentence classification task.

Many of the existing works use a limited set of labels: generic/non-generic or generic/specific. Herbelot and Copestake (2010, 2011) use a tagset similar to ours: ONE, SOME, MOST, ALL, QUANT. However, their concept is quite different: The authors assume that covertly quantified NPs are not generic but underspecified and label those NPs according to how many members of a class they refer to.

Bhakthavatsalam et al. (2020) create a large knowledge base of generic statements. The metadata in this knowledge base includes the term (restrictor) and the quantifier. They also include statements with an overt quantifier but "generic interpretation".

Contrasting the previously mentioned shallow annotation schemes, Bunt et al. (2018) / Bunt (2019) propose an annotation scheme for quantification that consists of several layers for syntactic and semantic representations but does not incorporate solutions for generics/habituals, yet.

Donatelli et al. (2019) suggest to expand the existing Abstract Meaning Representation (AMR) framework for the semantic annotation of sentences (Banarescu et al., 2013) by marking aspect and tense. As for aspect, they include such categories as "habitual" that characterise a regular recurrence of an event or state, and "stable" that characterises states and includes generalisations over kinds.

In comparison to the discussed approaches, we do not limit ourselves to NPs or clauses but annotate entire statements. Our tagset provides tags for overt quantification as well as covert generic quantification. Regarding the annotation of generalising statements, we jointly consider generalisations about entities, events and other types, that have been predominantly studied separately in the past.

8 Conclusion and Future Work

In this paper, we presented an annotation scheme for quantified phenomena using the example of generalisation. Our tagset matches every quantifier occurring in natural language to a particular tag, based on semantic criteria. We propose a shallow annotation that combines quantifier, restrictor and nuclear scope into a common annotation span, where the smallest unit of annotation is a clause. This approach is suitable for large-scale annotations which aim to investigate the distribution of quantified phenomena in a corpus, or to mark quantified statements to serve as feature input in follow-up applications. As a first step in this direction, we introduce our corpus of fictional texts that are annotated with generalising interpretations (among other phenomena).[8] Moreover, we are working on an automatic tagger for generalising statements.

Acknowledgements

Our special thanks go to Anke Holler and the other members of the MONA project group, who advised us on the development of this work. We further thank the anonymous reviewers for their valuable comments. This work is funded by Volkswagen Foundation (Dönicke, Gödeke), and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 424264086 (Varachkina).

[8] Our corpus and annotation guidelines are accessible at https://gitlab.gwdg.de/mona/korpus-public.

References

Jeremy Avigad and Richard Zach. 2020. The Epsilon Calculus. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2020 edition. Metaphysics Research Lab, Stanford University.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Maria Becker, Alexis Palmer, and Anette Frank. 2016. Argumentative texts and clause types. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 21–30, Berlin, Germany. Association for Computational Linguistics.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. GenericsKB: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*.

Harry Bunt. 2019. A semantic annotation scheme for quantification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 31–42, Gothenburg, Sweden. Association for Computational Linguistics.

Harry Bunt, James Pustejovsky, and Kiyong Lee. 2018. Towards an ISO standard for the annotation of quantification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Noel Burton-Roberts. 1976. On the generic indefinite article. *Language*, 52(2):427–448.

Gregory Carlson. 2011. Genericity. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, chapter 47, pages 1153–1185. De Gruyter Mouton.

Gennaro Chierchia and Sally McConnell-Ginet. 2000. *Meaning and grammar: An introduction to semantics*, 2nd edition. MIT press.

Ariel Cohen. 2004. Existential generics. *Linguistics and Philosophy*, 27(2):137–168.

Lucia Donatelli, Nathan Schneider, William Croft, and Michael Regan. 2019. Tense and aspect semantics for sentential AMR. *Proceedings of the Society for Computation in Linguistics*, 2(1):346–348.

Dudenredaktion. n.d. „manch“ auf Duden online. URL: https://www.duden.de/node/151861/revision/151897.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2003. *The measurement of interrater agreement*, 3rd edition, chapter 18. John Wiley & Sons.

Janet Dean Fodor and Ivan A Sag. 1982. Referential and quantificational indefinites. *Linguistics and philosophy*, 5(3):355–398.

Theodor Fontane. 1995. *The Stechlin*. Translated by William L. Zwiebel. Germ Series. Camden House. URL: https://books.google.de/books?id=7yh0ZJjNgxoC.

Theodor Fontane. 2012. Der Stechlin. In *TextGrid Repository*. Digitale Bibliothek. URL: https://hdl.handle.net/11858/00-1734-0000-0002-AECF-D.

Theodor Fontane. 2017. Der Stechlin. In Berenike Herrmann and Gerhard Lauer, editors, *KOLIMO. A corpus of Literary Modernism for comparative analysis*. Originally published in Fontane (2012).

Annemarie Friedrich and Alexis Palmer. 2014. Situation entity annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. 2015. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 21–30, Denver, Colorado, USA. Association for Computational Linguistics.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768.

Johann Wolfgang Goethe. 2012. Die Wahlverwandtschaften. In *TextGrid Repository*. Digitale Bibliothek. URL: https://hdl.handle.net/11858/00-1734-0000-0006-6A93-D.

Johann Wolfgang von Goethe. 19--? *Elective affinities : a novel*. URL: https://archive.org/details/electiveaffiniti00goetuoft/page/12/mode/2up.

Johann Wolfgang von Goethe. 2017. Die Wahlverwandtschaften. In Berenike Herrmann and Gerhard Lauer, editors, *KOLIMO. A corpus of Literary Modernism for comparative analysis*. Originally published in Goethe (2012).

Martin Hackl. 2009. On the grammar and processing of proportional quantifiers: *most* versus *more than half*. *Natural language semantics*, 17(1):63–98.

Irene Roswitha Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts.

Aurelie Herbelot and Ann Copestake. 2010. Annotating underquantification. In *Proceedings of the*

Fourth Linguistic Annotation Workshop, pages 73–81, Uppsala, Sweden. Association for Computational Linguistics.

Aurelie Herbelot and Ann Copestake. 2011. Formalising and specifying underquantification. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Berenike Herrmann and Gerhard Lauer. 2017. KOLIMO. A corpus of Literary Modernism for comparative analysis. URL: https://kolimo.uni-goettingen.de/about.

Klaus von Heusinger. 2000. The reference of indefinites. In Klaus von Heusinger and Urs Egli, editors, *Reference and anaphoric relations*, pages 247–265. Springer.

Hans Kamp. 1981. A theory of truth and semantic representation. In Jeroen A. G. Groenendijk, Theo M. V. Janssen, and Martin J. B. Stokhof, editors, *Formal Methods in the Study of Language (Part I)*, pages 277–322. Mathematisch Centrum, Amsterdam.

Edward Keenan and Denis Paperno. 2012. *Handbook of quantifiers in natural language*, volume 90. Springer.

Oriana Kilbourn-Ceron. 2014. Almost: scope and covert exhaustification. In *West Coast Conference on Formal Linguistics (WCCFL)*, volume 32, pages 121–130.

Manfred Krifka. 2016. Quantification and information structure. In *The Oxford handbook of information structure*.

Manfred Krifka and Claudia Gerstner. 1987. An outline of genericity. Seminar für natürlich-sprachliche Systeme der Universität Tübingen.

Manfred Krifka, Francis Jeffry Pelletier, Gregory Carlson, Alice ter Meulen, Gennaro Chierchia, and Godehard Link. 1995. Genericity: An introduction. In Greg N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, pages 1–124. University of Chicago Press.

Silke Lahn and Jan Christoph Meister. 2016. *Einführung in die Erzähltextanalyse*, 3 edition. J.B. Metzler, Stuttgart.

Sarah-Jane Leslie, Sangeet Khemlani, and Sam Glucksberg. 2011. Do all ducks lay eggs? the generic overgeneralization effect. *Journal of Memory and Language*, 65(1):15–31.

Sarah-Jane Leslie and Adam Lerner. 2016. Generic Generalizations. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, winter 2016 edition. Metaphysics Research Lab, Stanford University.

David Lewis. 1975. Adverbs of quantification. In Edward L. Keenan, editor, *Formal Semantics of Natural Language*. Cambridge University Press.

Sebastian Löbner. 2005. *Quantoren im GWDS*, pages 171–192. Max Niemeyer Verlag.

Louise McNally. 1998. Existential sentences without existential quantification. *Linguistics and Philosophy*, 21(4):353–392.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim. 2003. ACE-2 Version 1.0. *Linguistic Data Consortium, Philadelphia*. URL: https://catalog.ldc.upenn.edu/LDC2003T11.

Friederike Moltmann. 2013. The semantics of existence. *Linguistics and Philosophy*, 36(1):31–63.

Anna Nedoluzhko. 2013. Generic noun phrases and annotation of coreference and bridging relations in the Prague dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 103–111, Sofia, Bulgaria. Association for Computational Linguistics.

Barbara H. Partee. 1990. Domains of quantifications and semantic typology. Mid-America Linguistics Conference.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 40–49.

Laura Rimell. 2004. Habitual sentences and generic quantification. In *Proceedings of WCCFL*, volume 23, pages 663–676.

Achim Vesper. 2014. *Literatur und Aussagen über Allgemeines*, pages 181–196. De Gruyter (A).

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Yoad Winter. 1997. Choice functions and the scopal semantics of indefinites. *Linguistics and Philosophy*, 20(4):399–467.

Edward N. Zalta. 2020. Gottlob Frege. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2020 edition. Metaphysics Research Lab, Stanford University.

A Examples for Quantified Statements with Generalising Interpretations

(46) [In all natural objects with which we are acquainted, we observe immediately that they have a certain relation to themselves]$_{\text{ALL}}$. (Goethe, EA, p. 53)

(47) [Heroism is the exception]$_{\text{BARE}}$ [and mostly the product of a separate situation]$_{\text{MEIST}}$. (Fontane, Stechlin, p. 19)

(48) [But there are also fruits which are not outward, which are of the true germinal sort, and which develop themselves sooner or later in a beautiful life]$_{\text{EXIST}}$. (Goethe, EA, p. 43)

(49) On New Year's Day he was to follow him, and spend the Carnival at his house in the city, where Luciana was promising herself infinite happiness from a repetition of her charmingly successful pictures, [as well as from a hundred other things]$_{\text{ZAHL}}$; (Goethe, EA, p. 240)

(50) In providing against accidents, [which, though common, yet only too often find us unprepared]$_{\text{DIV}}$, they thought it especially necessary to have at hand whatever is required for the recovery of drowning men (Goethe, EA, p. 48)

(51) "[We are strange creatures]$_{\text{BARE}}$," said Edward, smiling. "[If we can only put out of sight anything which troubles us, we fancy at once we have got rid of it]$_{\text{BARE}}$. (Goethe, EA, p. 25)

(52) [A fellow from Friesack better not have a name like Raoul]$_{\text{NEG}}$. (Fontane, Stechlin, p. 4)

B German Versions of Translated Examples

Numbers are identical to those of the corresponding translations in the main part of the paper. Note that the German version sometimes receives another annotation than its English translation because the quantification may differ.

(8') [...] [und augenblickliche Hülfe ist doch immer das, was auf dem Lande am meisten vermißt wird]$_{\text{ALL}}$. (Goethe, WV)

(9') [...] trenne alles, was eigentlich Geschäft ist, vom Leben! (Goethe, WV)[9]

(13') Gedenken wir nur des Kalks, [der zu allen Säuren eine große Neigung, eine entschiedene Vereinigungslust äußert]$_{\text{ALL}}$! (Goethe, WV)

(14') [Die Männer denken mehr auf das Einzelne, auf das Gegenwärtige]$_{\text{BARE}}$ [...] (Goethe, WV)

(16') [...] [da sind zum dritten die, die den Meeresgrund absuchen wie 'ne Wiese]$_{\text{EXIST}}$ (Fontane, Stechlin)

(17') [Aber es gibt dergleichen noch, es muß dergleichen geben oder doch wieder geben]$_{\text{EXIST}}$. (Fontane, Stechlin)

(25') Die Grafschaft habe immer konservativ gewählt; es sei Ehrensache, wieder konservativ zu wählen. »[Und ob die Welt voll Teufel wär']$_{\text{BARE}}$. (Fontane, Stechlin)

(26') Da kommt hier so Anno Domini ein Burggraf ins Land, und das Land will ihn nicht, [und er muß sich alles erst erobern, die Städte beinah und die Schlösser gewiß]$_{\text{ALL}}$. (Fontane, Stechlin)

(28') Wenn es mir erlaubt ist, dem Scheine nach weit auszuholen, so sind wir bald am Platze. (Goethe, WV)

(29') [Unsere vortreffliche Vorsteherin läßt mich gewöhnlich die Briefe lesen, in welchen sie Beobachtungen über ihre Zöglinge den Eltern und Vorgesetzten mitteilt]$_{\text{DIV}}$. (Goethe, WV)

(30') »Es betrifft unsern Freund, den Hauptmann,« antwortete Eduard. »Du kennst die traurige Lage, [in die er, wie so mancher andere, ohne sein Verschulden gesetzt ist]$_{\text{DIV}}$. (Goethe, WV)

(31') Dieser sein Eintritt ins Regiment fiel so ziemlich mit dem Regierungsantritt Friedrich Wilhelms IV. zusammen, [und wenn er dessen erwähnte, so hob er, sich selbst persiflierend, gerne hervor, »[daß alles Große seine Begleiterscheinungen habe]$_{\text{ALL}}$«]$_{\text{BARE}}$. (Fontane, Stechlin)

(39') [Die Landleute haben die rechten Kenntnisse]$_{\text{BARE}}$; [ihre Mitteilungen aber sind konfus]$_{\text{BARE}}$ [und nicht ehrlich]$_{\text{NEG}}$. [Die Studierten aus der Stadt und von den Akademien sind wohl klar und ordentlich]$_{\text{BARE}}$, [aber es fehlt an der unmittelbaren Einsicht in die Sache]$_{\text{BARE}}$. (Goethe, WV)

(43') [Und doch ist es in manchen Fällen [...] notwendig und freundlich]$_{\text{DIV}}$,

[9] In English, this sentence is declarative, whereas in German it is imperative. Therefore, it is not annotated in German.

[lieber nichts zu schreiben, als nicht zu schreiben]$_{\text{NEG}}$. (Goethe, WV)

(46') [An allen Naturwesen, die wir gewahr werden, bemerken wir zuerst, daß sie einen Bezug auf sich selbst haben]$_{\text{ALL}}$. (Goethe, WV)

(47') [Heldentum ist Ausnahmezustand]$_{\text{BARE}}$ [und meist Produkt einer Zwangslage]$_{\text{MEIST}}$. (Fontane, Stechlin)

(48') [...] [aber es gibt auch verschlossene Früchte]$_{\text{EXIST}}$, [die erst die rechten, kernhaften sind und die sich früher oder später zu einem schönen Leben entwickeln]$_{\text{BARE}}$. (Goethe, WV)

(49') Aufs Neujahr sollte ihm dieser folgen und das Karneval mit ihm in der Stadt zubringen, [wo Luciane sich von der Wiederholung der so schön eingerichteten Gemälde sowie von hundert andern Dingen die größte Glückseligkeit versprach]$_{\text{ZAHL}}$ (Goethe, WV)

(50') [Da man auch die gewöhnlichen und dessen ungeachtet nur zu oft überraschenden Notfälle durchdachte, so wurde alles, was zur Rettung der Ertrunkenen nötig sein möchte, um so mehr angeschafft]$_{\text{DIV}}$ (Goethe, WV)

(51') »[Wir sind wunderliche Menschen]$_{\text{BARE}}$,« sagte Eduard lächelnd. »[Wenn wir nur etwas, das uns Sorge macht, aus unserer Gegenwart verbannen können, da glauben wir schon, nun sei es abgetan]$_{\text{BARE}}$. (Goethe, WV)

(52') [Wer aus Friesack is, darf nicht Raoul heißen]$_{\text{NEG}}$. (Fontane, Stechlin)

The ISA-17 Quantification Challenge: Background and introduction

Harry Bunt
Department of Cognitive Science and Artificial Intelligence
School of Humanities and Digital Sciences
Tilburg University, Tilburg, Netherlands
bunt@tilburguniversity.edu

Abstract

This short paper provides background information for the shared quantification annotation task at the ISA-17 workshop, a.k.a. the Quantification Challenge. The role of the abstract and concrete syntax of the QuantML markup language are explained, and the semantic interpretation of QuantML annotations in relation to the ISO principles of semantic annotation. Additionally, the choice of the test suite of the Quantification Challenge is motivated.

1 Introduction

The ISA-17 Quantification Challenge was motivated by the decision of the International Organisation for Standardisation ISO to develop an international standard for the annotation of quantification in natural language, extending the series of standards for semantic annotation called the ISO Semantic Annotation Framework (SemAF, ISO 24617)). Other parts of this series include standards for the annotation of

(1) time and events (ISO 24617-1, 'ISO-TimeML');

(2) dialogue acts (ISO 24617-2, 'DiAML');

(3) semantic roles (ISO 24617-4);

(4) spatial information (ISO 24617-7, 'ISO-Space');

(5) discourse relations (ISO 24617-8, 'DR-Core');

(6) coreference (ISO 24617-9, 'Reference Annotation Framework').

Also belonging to this series is the meta-standard ISO 24617-6, 'Principles of semantic annotation', which defines a common methodological framework for developing other parts of SemAF.

As the first steps in the development of an annotation standard for quantification, preliminary studies have been conducted and reported at LREC 2018 (Bunt, Lee and Pustejovsky, 2018), at IWCS 2019 (Bunt, 2019), and in a technical report of Tilburg University (Bunt, 2021) in which the markup language QuantML is defined. On the basis of these studies, the document ISO WD 24617-12 was drafted. The ISA-17 Quantification Challenge is intended to identify the strengths, limitations, and deficiencies of the QuantML proposal by inviting experts in quantification and/or in semantic annotation to explore the application of QuantML in the annotation of a range of test sentences that display some of the phenomena that the future standard would hope to cover.

This paper is organised as follows. Section 2 briefly summarizes the ISO Principles of semantic annotation, especially where it concerns the architecture of a semantic annotation scheme, including the design of an abstract syntax and the specification of a concrete syntax plus the significance of a compositional semantics of the (abstract syntactic structures of the) annotations, and applies this to the QuantML annotation scheme. Section 3 introduces and motivates the choices in the test suite used in the Quantification Challenge.

2 QuantML

2.1 Annotation scheme architecture

The usual definition of a markup language consists of the specification of a number of XML elements, attributes, and values, that can be used to form descriptions of the linguistic properties of certain stretches of text or speech, called 'markables'. The definitions of TimeML (Pustejovsky et al., 2007) and SpatialML (Mani et al, 2010) illustrate this.

According to the ISO Principles of semantic annotation ISO 24617-6; see also Bunt (2015) and

Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pages 33–40
June 16–17, 2021.

Pustejovsky et al. (2017)) a semantic annotation scheme has a three-part architecture consisting of (1) an abstract syntax that specifies the possible *annotation structures* at a conceptual level as set-theoretical constructs, such as pairs and triples of concepts; (2) a concrete syntax, that specifies a representation format for annotation structures (for example using XML); (3) a semantics that specifies the meaning of the annotation structures defined by the abstract syntax.

The distinction of an abstract and a concrete syntax is motivated by the fundamental distinction between 'annotations' and 'representations', made in the Linguistic Annotation Framework (ISO standard 24612, see also Ide and Romary (2004). An 'annotation' captures certain linguistic information, independent on a particular representation format, while a 'representation' specifies a format for representing annotations. In the three-part architecture, 'annotations' are the conceptual structures defined by the abstract syntax (and called 'annotation structures'); 'representations' correspond to the particular format in which these structures are expressed (which we will usually call 'annotations', following the most common usage of this term). ISO standards for semantic annotation are intended to apply not at the level of representation formats, but that of the information they represent: the level of conceptual annotation structures.

The third component of a semantic annotation scheme, the specification of a semantics of annotation structures, is a requirement specific of *semantic* annotations, the *requirement of semantic adequacy* (Bunt and Romary (2002)): if the annotations would not have a well-defined semantics, it would not be clear what semantic information they add to the natural language expressions they annotate. Defining the semantics at the level of the *abstract* syntax puts the focus of an annotation standard at the conceptual level, rather than at the level of representation formats.

Formally, the definition of an annotation scheme is a triple consisting of specifications of an abstract syntax (AS), a concrete syntax (CS), and a semantics ($ASem$):

(1) $A = \langle ASyn_a, ASyn_c, ASem \rangle$

The abstract syntax consists of the specification of a set of basic concepts, called the 'conceptual inventory' (CI), and a set of constructions (AC) for forming conceptual structures out of basic concepts.

(2) $ASyn_a = \langle CI, AC \rangle$

Together, the sets CI and AC define the class of well-formed annotation structures.

The concrete syntax specification $ASyn_C$ contains a vocabulary V_c, the specification CC of a class of syntactic structures, such as XML elements, and an encoding function F_e. The components V_c and CC together define a class of well-formed representations, and F_e assigns such a representation to every well-formed annotation structure.

(3) $ASyn_c = \langle V_c, CC, F_e \rangle$

The semantics $ASem$ can be specified in various ways, for example as a model-theoretic semantics $\langle M, I_M \rangle$ with a model M and an interpretation function I_M that assigns concepts from M as meanings to annotation structures.

The three parts of the annotation schema are related through the encoding function F_e and the interpretation function I_M. In particular, a requirement for the relation between abstract and concrete syntax is that the concrete syntax is *complete* and *unambiguous* (Bunt (2010)) for the abstract syntax, i.e. every annotation structure has a representation in the concrete syntax, and every representation is the encoding of exactly one annotation structure. In other words, F_e is a total function and so is its inverse F_e^{-1}. The semantic component should also be complete: every annotation structure has a semantic interpretation.

Two types of structure are distinguished in an abstract syntax: *entity structures* and *link structures*. An entity structure contains semantic information about a segment of primary data and is formally a pair $\langle m, s \rangle$ consisting of a markable (m), which refers to a segment of primary data, and certain semantic information (s). A link structure contains information about the way two or more segments of primary data are semantically related. In QuantML three types of entity structure are defined (participant structures, event structures, and modifier structures) and two types of link structure (participation links and scope links). Participation links relate participants to events; scope links indicate scope relations between participants. See further Section 2.3.

The three-part structure of a semantic annotation scheme does not need to frighten the users of such a scheme: annotators (human or automatic) only have to deal with concrete representations. They can rely, however, on the abstract syntax and its

semantics that comes with the definition of the scheme, in particular when in doubt how to use the concrete syntax for annotating certain linguistic phenomena: rather than just relying on annotation guidelines, which are bound to be incomplete, they check the semantics for the precise implications of the choices offered by the concrete syntax.

2.2 QuantML concrete syntax

A concrete QuantML syntax is specified here in the form of an XML representation of annotation structures. For each type of entity structure, defined by the abstract syntax, a corresponding XML element is defined; each of these elements has an attribute @xml:id whose value is a unique identification of (the information in) the element, and an attribute @target, whose value anchors the annotation in the primary data, having a markable as value (or a sequence of markables). In addition, these elements have the following attributes:

1. the XML element <entity>, for representing participant structures, has the attributes @domain, @involvement, @definiteness and optionally @size (default value: ≥ 1;

2. the XML element <event>, for representing event structures, has the attribute @pred for specifying an event type;

3. the XML element <qDomain>, for representing a quantification domain: has the attributes @source (with multiple values in the case of a conjunctive specification) and @restrictions;

4. the XML element <sourceDomain>, for representing quantification source domain specifications without modifiers: has the attributes @pred and @individuation;

5. the XML element <adjMod>, for representing adjectival modifiers, with the attributes @pred and @distr, and optionally the attribute @restrictions;

6. the XML element <nnMod>, for representing nouns as modifiers, with the attributes @pred and @distr, and optionally @restrictions;

7. the XML element <ppMod>, for representing PP modifiers, with the attributes @pRel, @pEntity, @distr and @linking;

8. the XML element <relClause>, for representing relative clauses, with the attributes @semRole, @clause, @distr and @linking;

9. the XML element <possRestr>, for representing possessive restrictions, with the attributes @possessor, @distr, and @linking.

For the two types of link structure defined by the abstract syntax, a corresponding XML element is defined:

- <participation> has the attributes @event, @participant, @semRole, @distr, and @evScope (default value: "narrow"), and optionally @exhaustiveness (default value: "non-exhaustive"), @rep (repetitiveness, default value: ≥ 1), and @polarity (default value: "positive");

- <scoping> has the attributes @arg1, @arg2, @scopeRel.

2.3 QuantML abstract syntax

The QuantML abstract syntax defines the following entity structures $\langle m, s\rangle$ with markable m and semantic content s:

1. Participant structures: s is a triple or quadruple $\langle DS, q, d, N\rangle$, where DS is a domain specification, q is a specification of domain involvement, d is a definiteness, and N is a numerical size specification (optional).

2. Event structures: s is a predicate denoting an event domain.

3. Modifier structures: s contains a predicate for (NP head) noun modification by an adjectives, noun, prepositional phrase, relative clause, or possessive restriction, plus parameters for specifying properties of the modification.

The following link structures are defined:

1. Participation links: A 6-9 tuple as shown in (4), where the first two components are the linked event and participant structures, and the other components indicate properties of the way in which the participants are involved in the events, specifying a semantic role (R), a distribution (d), an event scope (σ) that specifies whether the event structure has wider or narrower scope than the participant structure, and optionally an exhaustiveness (ξ), a repetitiveness (ρ), and a polarity (p).

(4) $L_{P1} = \langle \epsilon_e, \epsilon_p, R, d, \sigma, p \rangle$

2. Scope relation links: triples ⟨participation link, participation link, scope relation⟩.

The conceptual inventory of the abstract syntax includes:[1]

1. predicates that characterise quantification domains, corresponding to the meanings of common nouns of the language of the primary data;

2. predicates that characterise event domains, corresponding to the meanings of verbs (and some other lexical items);

3. predicates corresponding to the meanings of adjectives or prepositions;

4. relations that denote semantic roles; for this purpose, the semantic roles defined in ISO 245617-4 (Semantic roles) are used;

5. binary and ternary relations for specifying proportional domain involvement, such as most, 'half, 'total', and "between";

6. non-numerical quantitative predicates for specifying domain involvement, like some and several;

7. parameters for specifying definiteness, polarity, distributivity, individuation, relative scoping, repetitiveness and exhaustivity.

Quantification annotation is associated with the units that in linguistics are called (small) clauses, i.e. a finite verb and its arguments. This is the level of syntactic structure where issues arise of the relative scoping of quantified participants in different roles, as well as relative scoping of event quantification and participant quantification. Annotation structures at this level are quadruples consisting of an event structure, a set of participant structures, a set of link structures that relate participants to events, and a set of link structures that specify scope relations; see (5), where ϵ_{ev} is an event structure; $\epsilon_{P1}...\epsilon_{Pn}$ are participant structures; L_{P1}, ..., L_{Pn} are participation link structures, and $sc_1, ..., sc_k$ are scope link structures.

(5) $A = \langle \epsilon_{ev}, \{\epsilon_{P1}, ..., \epsilon_{Pn}\}, \{L_{P1}, ..., L_{Pn}\}, \{sc_1, ..., sc_k\} \rangle$

[1] This listing is slightly simplified. For the full specification see Bunt (2020).

2.4 Semantics

The design of QuantML was inspired by the theory of generalized quantifiers (GQT, Barwise and Cooper 1981; Keenan and Westerstaahl), 1997, combined with neo-Davidsonian event semantics (Davidson, 1967; Parsons, 1990), viewing natural language quantifiers as properties of sets of participants involved in sets of events. Champollion (2015) has shown the viability of this type of combination.

QuantML has an interpretation-by-translation semantics in the form of a compositional, recursive translation of annotation structures to Discourse Representation Structures (DRSs) as defined by (Kamp and Reyle, 1993). If the annotation structure is fully connected, i.e., if (1) all participant entity structures are linked to an event structure, and (2) for any two participant entity structures linked to the same event structure the relative scopes are specified, then the interpretation function delivers a standard DRS; if one or both of these conditions are not satisfied, then the interpretation delivers an underspecified DRS (UDRS, Reyle, 1984).

The QuantML semantics is compositional in the sense that the interpretation of an annotation structure is obtained by combining the interpretations of its component entity structures and participation link structures in a manner that is determined by the scope link structures. Combining GQT Casting the semantics in this form is particularly convenient for combining annotations of quantification with other types of semantic information, using annotation schemes of the ISO Semantic Annotation Framework (SemAF) and annotation scheme plug-ins (Bunt, 2019).

The semantic entities that correspond to participant entity structures may be of any kind: real-world objects, abstract entities, events, individual concepts, intentional and intensional entities, hypothetical and fictional entities. The design of QuantML aims to be neutral with respect to ontological and linguistic views on the existence of objects of various kinds and the need for them in semantic accounts of natural language.

Note that a participation link structure embeds the linked event structure and participant structure, to the effect that the annotation structures as defined by the abstract syntax are nested structures, as opposed to their flat XML-representations. The interpretation of a fully-connected annotation structure is therefore determined by the interpretation

of the participation link structures.

The semantics of a participation link structure is a combination of the semantics of its components by means of the interpretation function I_Q as specified in (6), where $\cup$ is the operation of merging two DRSs, as defined in DRT, and $\cup^*$ is the scoped merge operation, defined below in (11).

(6) a. $I_Q(\epsilon_E, \epsilon_P, R,$ d, narrow) = $(I_Q(\epsilon_P) \cup^* I_Q(\epsilon_E)) \cup I_Q$(R, d, narrow)

b. $I_Q(\epsilon_E, \epsilon_P, R,$ d, wide) = $(I_Q(\epsilon_E) \cup^* I_Q(\epsilon_P)) \cup I_Q$(R, d, wide)

c. $I_Q(\epsilon_E, \epsilon_P, R,$ d, free) = $(I_Q(\epsilon_E) \cup I_Q(\epsilon_P)) \cup I_Q$(R, d, free)

As an illustration, consider sentence (7), with its annotation and interpretation shown in Figure 1.

(7) All the students read three papers

A quantifier of the form *"All the D"* is interpreted as a DRS of the form (8), where capital letters are used for discourse referents that correspond to non-empty sets of individuals. This DRS says that there is a non-empty subset X of the quantification domain D containing all the contextually distinguished students, using the subscript '0' to indicate the contextually determined 'reference domain' or 'context set' (Westerståhl, 1985)). This subset X contains those elements of the reference domain that participate in a set of events. For the quantifier *"All the students"* this leads to the interpretation (8b). Similarly, the annotation of the quantifier *"three papers"* leads to the interpretation (8c).

(8) a. $[X|x \in X \leftrightarrow D_0(x)]$

b. $[X|x \in X \leftrightarrow \text{student}_0(x)]$

c. $[Y||Y| = 3, y \in Y \rightarrow paper(\text{y})\ ]$

For the semantic role R, the distribution d = 'individual', and the event scope σ = 'narrow', the interpretation of the third component in (6) is the DRS in (9), which says that there is a non-empty participant set of which every member has the role R in a non-empty set of events:

(9) I_Q(R,individual,narrow) = $[X|x \in X \leftrightarrow D_0(x), x \in X \rightarrow [E|e \in E \rightarrow R(e,x)]]$

Application of (6) and merging the DRS in (9) with the DRSs interpreting the participant structure and the event structure, results in (10) for the interpretation of the annotation of the sentence in (7).

(10) $[X|x \in X \leftrightarrow \text{student}_o(x), x \in X \rightarrow [Y||Y| = 3, y \in Y \rightarrow [E|\text{paper}(y), e \in E \rightarrow [|agent(e,x), theme(e,y)]]]]$

The scoped merge operation is designed to combine the information about quantified participation in two participation link structures, and is defined as follows:

(11) The scoped merge operation combines the information in its argument DRSs into a DRS that reflects the relative scoping of the quantifications involved, as well as the relative scopings of participants and events, while unifying the event discourse referents in the two arguments. (If this unification is not possible, then the operation fails.)

For annotation structures that do not fully specify the relative scopes of all the sets of participants involved in the same events, the semantic interpretation takes the form of a set of (sub-)DRSs that express the semantics of the participation link structures, plus the scope restrictions for their possible combination. Such an interpretation is known in DRT as an underspecified DRS (UDRS, Reyle, 1994).

A detailed specification of the semantics of QuantML annotation structures can be found in the technical report Bunt (2020), available on the ISA-17 website https://sigsem.uvt.nl/isa17/TiCC_Report_Quantification-12-Print.pdf.

3 The Quantification Challenge test suite

3.1 Quantification phenomena

The Quantification Challenge test suite has been constructed in such a way that its sentences illustrate the coverage of the QuantML proposal, with a number of challenging borderline cases that invite speculation and creativeness in finding adequate annotations. More specifically, the test suite covers the following phenomena:

- Definiteness and determinacy of NPs. Where an NP like *"the students"* is obviously definite, and semantically determinate, less obvious is how to characterize *"some of the students"* or *"one of my friends"*.
- Attributive and predicative adjectives.
- Deictic NPs such as *"I"* and *"you"*.

(7) All the students read three papers.

Markables:
m1 = all the students, m2 = students, m3 = read, m4 = three papers, m5 = papers.

QuantML annotation:
<entity xml:id=x1 target=#m1 domain=#x2 involvement=all definiteness=det/>
<sourceDomain xml:id=x2 target=#m2 pred=student/>
<event xml:id=e1 target=#m3 pred=read/>
<entity xml:id=x3 target=#m4 domain=#x4 involvement=3 definiteness=indet/>
<sourceDomain xml:id=x4 target=#m5 pred=paper/>
<participation event=#e1 participant=#x1 semRole=agent distr=individual evScope=narrow/>
<participation event=#e1 participant=#x3 semRole=theme distr=individual evScope=narrow/>
<scoping arg1=#x1 arg2=#x2 scopeRel=wider/>

Annotation structure:
$A = \langle \epsilon_{ev}, \{\epsilon_{P1}, \epsilon_{P2}\}, \{L_{P1}, L_{P2}\}, \{sc_1\}\rangle =$
$= \langle\langle$m3, read$\rangle, \{\langle m1, \langle m2, \langle$student,count$\rangle\rangle$, all, det$\rangle, \langle m4, \langle m5, \langle$paper,count$\rangle\rangle, 3, indet\rangle\}$,
$\{\langle\langle m3, read\rangle, \langle m1, \langle m2, \langle student, count\rangle\rangle, all, indet\rangle\rangle, agent, individual, narrow\rangle$,
$\{\langle\langle m3, read\rangle, \langle m4, \langle m5, \langle paper, count\rangle\rangle, 3, indet\rangle\rangle, theme, individual, narrow\rangle\}$
$\langle\langle\langle$m3, read$\rangle, \{\langle m1, \langle m2, \langle$student,count$\rangle\rangle$, all, det$\rangle, \langle m4, \langle m5, \langle$paper,count$\rangle\rangle, 3, indet\rangle$, wider$\rangle\rangle$

c. Semantics:
$I_Q(A) = I_Q(L_{P1}) \cup^* I_Q(L_{P2}) \cup^* I_Q)(\epsilon_{ev}) =$
$[X|x \in X \leftrightarrow$ student$_0(x), x \in X \rightarrow [Y|y \in Y \rightarrow [E|$ paper$(y), e \in E \rightarrow [$ agent(e, x), theme$(e, y)]]]]]$

Figure 1: Example annotation with abstract syntax and semantics

- Scope ambiguities, as in *"The editors didn't see a misprint"*.
- Conjoined NPs, like *"Bert and Alice"*.
- Relative clauses.
- Proper names.
- Temporal quantifiers, such as *"twice", "two to three times"*, and *"every hour"*.
- Negations.
- Mass NPs.
- Anaphoric possessive pronouns (*"his", "their"*).
- Complex possessives, as in *"The headmaster's childrens' toys"*.
- Collective quantifications, as in *"These machines combine 12 parts"*, interpreted as saying that each of the machines every time combines twelve parts.
- Exhasutive quantification.
- Quantification with unscpecific distribution, as in *"The boys carried the boxes upstairs"*.
- Complex NN-modifications, like *"new corona virus infections"*.

3.2 Markables

The sentences in the test suite all come with a suggestion for substrings to be used as markables in the annotation. This is to make the comparison of annotations made by different annotators easier, .

Concerning the choice of markables for a given (small) clause, first of all every NP is naturally a markable, describing a set of participants (or possibly a single participant), and the main verb (possibly with modifiers) is another markable, corresponding to the events in which the participants are involved. Other markables are those words that correspond to the predicates of the conceptual inventory in the abstract syntax and those in the concrete syntax, notably as values of the @pred attribute. This concerns all nouns, adjectives, prepositions, and numerical as well as non-numerical terms.

A direct consequence of this way of distinguish-

ing markables, is that they may overlap; for example, the markable for an NP overlaps with the one for its head noun. In such a case, the numbering of markables is determined in the first place by its left boundary, and if they have the same left boundary, than by the linear position of the right boundary. So in the sentence *"Most of the students passed the exam"*, *Most"* is numbered as markable m1, and *"Most of the students"* as markable m2.

Markables may be discontinuous, for example, in *"The boys carried the boxes upstairs"*, the words *"carried upstairs"* is a discontinuous markable. Following the ordering convention of Discontinuous Phrase Structure Grammar (Bunt, 1996), the numbering of discontinuous markables is determined by their leftmost element, and if two such markables have the same leftmost element then by their next element, and so on.

3.3 Annotation guidelines

The documentation for annotating and interpreting the sentences of the test suite, in particular the technical report (Bunt, 2018), defines concepts and provides guidelines for dealing with these phenomena. These guidelines have not yet been very well developed, and a secondary purpose of the ISA-17 Quantification Challenge, besides the identification of its strengths and weaknesses for annotating quantification phenomena in a semantically adequate way, is to obtain a good picture of the ways in which these guidelines can be improved and extended. The introduction of decision trees to support annotators in choosing the right values of QuantML attributes may for example be an attractive direction.

References

J. Barwise and R. Cooper. 1981. Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, 4:159–219.

H. Bunt. 1996. Formal tools for the description and processing of discontinuous constituents. In Harry Bunt and Arthur van Horck, editors, *Discontinuous Constituency.*, pages 63–85. Mouton De Gruyter, Berlin.

H. Bunt. 2009. The DIT++ taxonomy for functional dialogue markup. In *Proceedings of AAMAS 2009 Workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts'*, pages 13–24, Budapest.

H. Bunt. 2010. A methodology for designing semantic annotation languages exploring semantic-syntactic ISO-morphisms. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, pages 29–46, Hong Kong: City University.

H. Bunt. 2015. On the principles of semantic annotation. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, pages 1–13, London.

H. Bunt. 2018. *Semantic Annotation of Quantification in Natural Language. TiCC Technical Report* 2018-15. Tilburg Center for Cognition and Communication, Tilburg University.

H. Bunt. 2019a. An annotation scheme for quantification. In *Proceedings 13th International Conference on Computational Semantics (IWCS 2019)*, pages 31–43, University of Gothenburg.

H. Bunt. 2019b. Plug-ins for content annotation of dialogue acts. In *Proceedings 15th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-15)*, pages 34–45, Gothenburg, Sweden.

H. Bunt. 2021. *Semantic Annotation of Quantification in Natural Language, second edition. TiCC Technical Report* 2021-2. Tilburg Center for Cognition and Communication, Tilburg University.

H. Bunt, J. Pustejovsky, and K. Lee. 2018. Towards an ISO Standar for the Annotation of Quantification. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

H. Bunt and L. Romary. 2002. Towards Multimodal Content Representation. In *Proceedings of LREC 2002, Workshop on International Standards of Terminology and Language Resources Management*, pages 54–60, Las Palmas. Paris: ELRA.

L. Champollion. 2015. The interaction of compositional semantics and event semantics. *Linguistics and Philosophy*, 38 (1):31–66.

D. Davidson. 1967. The logical form of action sentences. In N. Rescher, editor, *The Logic of Decision and Action*. University of Pittsburgh Press.

N. Ide and L. Romary. 2004. International standard for a linguistic annotation framework. *Natural Language Engineering*, 10:221–225.

ISO. 2010. *ISO 24612: Language resource management: Linguistic annotation framework (LAF)*. International Organisation for Standardisation ISO, Geneva.

ISO. 2012. *ISO 24617-1: Language resource management – Semantic annotation framework – Part 1: Time and events ('ISO-TimeML')*. International Organisation for Standardisation ISO, Geneva.

ISO. 2014. *ISO 24617-2: Language resource management – Semantic annotation framework – Part 4: Semantic roles*. International Organisation for Standardisation ISO, Geneva.

ISO. 2016a. *ISO 24617-2: Language resource management – Semantic annotation framework – Part 6:Principles of semantic annotation*. International Organisation for Standardisation ISO, Geneva.

ISO. 2016b. *ISO WD 24617-2: Language resource management – Semantic annotation framework – Part 12: Quantification ('QuantML', Working Draft)*. International Organisation for Standardisation ISO, Geneva.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht.

E. Keenan and D. Westerståhl. 1997. Generalized Quantifiers in Linguistics and Logic. In *Generalized Quantifiers in Natural Language*, pages 837–993. Foris, Dordrecht.

I. Mani, C. Doran, D. Harris, J.Hitzeman, S.R. Quinby, J. Richer, B. Wellner, S. Mardis, and S. Clancy. 2010. SpatialML: Annotation Scheme, Resources, and Evaluation. *Language Resources and Evaluation*, 44(3):263–280.

T. Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press, Cambridge, MA.

S. Peters and D. Westerståhl. 2013. The Semantics of Possessives. *Language*, 89(4):713–759.

J. Pustejovsky, H. Bunt, and A. Zaenen. 2017. Designing annotation schemes: From theory to model. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 21–72. Springer, Berlin.

J. Pustejovsky, R. Knippen, J. Litman, and R. Sauri. 2007. Temporal and event information in natural language text. In *Computing Meaning, Vol. 3*, pages 301–346. Springer, Dordrecht.

U. Reyle. 1993. Dealing with ambiguities by underspecification: Construction, representation, and deduction. *Journal of Semantics*, pages 123–179.

D. Westerståhl. 1985. Determiners and context sets. In Johan van Benthem and Alice ter Meulen, editors, *Generalized Quantifiers in Natural Language*, pages 45–71. Foris, Dordrecht.

Discourse-based Argument Segmentation and Annotation

Ekaterina Saveleva and **Volha Petukhova** and **Marius Mosbach** and **Dietrich Klakow**
Spoken Language Systems Group, Saarland Informatics Campus
Saarland University, Saarbrücken, Germany
{esaveleva,vpetukhova,mmosbach,dklakow}@lsv.uni-saarland.de

Abstract

The paper presents a discourse-based approach to the analysis of argumentative texts based on the assumption that the coherence of a text should capture argumentation structure. Therefore, existing discourse analysis tools can be successfully applied for argument segmentation and annotation tasks. We tested widely used Penn Discourse Tree Bank parser (Lin et al., 2010) and the state-of-the-art neural network NeuralEDUSeg (Wang et al., 2018) and XLNet (Yang et al., 2019) models on discourse segmentation and discourse relation recognition tasks. The two-stage approach outperformed the PDTB parser by broad margin, i.e. the best achieved F1 scores of 21.2% for PDTB parser vs 66.37% for NeuralEDUSeg and XLNet models. Neural network models were fine-tuned and evaluated on the argumentative corpus showing a promising accuracy of 60.22%. The complete argument structures were reconstructed for further argumentation mining tasks. The reference Dagstuhl argumentative corpus containing 2,222 elementary discourse unit pairs annotated with the top-level and fine-grained PDTB relations will be released to the research community.

1 Introduction

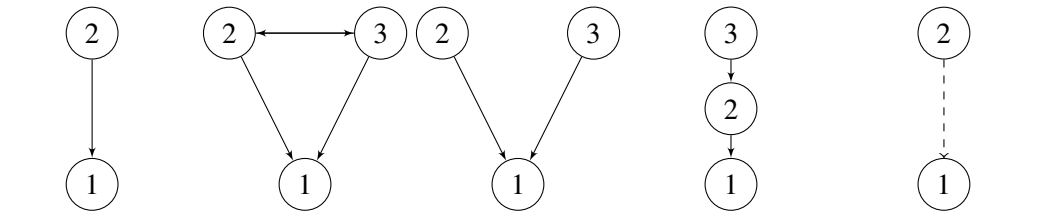
Enormous and ever growing digital content provides information where opinions, sentiment and arguments can be identified and analysed. For example, news and social media content is searched to filter or weight the validity of statements (Rowe and Butters, 2009), to identify the presence of fake news and false claims (Popat et al., 2018), to analyse opinions in public discussions (Murakami and Raymond, 2010), to detect opinion manipulation (Cambria et al., 2010), to predict consumers sentiment (Bai, 2011), to study citizen engagement (Purpura et al., 2008), and to recognize stance in political online debates (Somasundaran and Wiebe, 2010; Walker et al., 2012). Arguments from legal (Moens et al., 2007), financial (Hogenboom et al., 2010) or medical (Sanchez Graillet and Cimiano, 2019) documents are extracted to support professional decision-making. Natural argumentation is the focus of numerous educational scenarios assessing student's essays quality (Stab and Gurevych, 2017) and training argumentation and debate skills (Ashley et al., 2007; Petukhova et al., 2017). Automatic extraction and analysis of arguments from heterogeneous data is one of the important tasks of *argumentation mining* which aims to provide structured data for computational models of argument and reasoning engines (Lippi and Torroni, 2016).

While for some applications, an argument can be considered as an atomic entity without internal structure, for others defining its structure becomes crucial. For example, to recognize the speaker 'stance'[1] in online debates, the whole post can be acknowledged as an argument in 'favour' or 'against' a certain motion. An argument is, therefore, analysed given the other supporting or attacking arguments (Dung, 1995). Other argumentation mining tasks require structured argumentation models, e.g. tasks that aim at understanding and emulation of human inference, investigating patterns of reasoning, and tasks that focus on extraction and validity assessment of arguments.

Identification and classification of argument components are rather challenging tasks (Aharoni et al., 2014). The argument definition, the description of elementary units and building blocks of an argument, relations between and inside these units, the argument structures and argumentation schemes are still under debate. A simple argument structure is often considered as consisting of *a claim* that is supported by *evidence* (Mochales and Moens, 2011; Aharoni et al., 2014). A claim is an assertion that the argument aims to prove, i.e. a claim is a *conclusion* whose merit must be estab-

[1] Stance is defined as an overall position held by a person towards an idea or attitude (Somasundaran and Wiebe, 2009).

Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pages 41–53
June 16–17, 2021.



(i) basic argument (ii) linked support (iii) multiple support (iv) serial support (v) example support

Figure 1: Basic support relations and complex formations suggested by Peldszus and Stede (2013).

lished. Evidence presents (a set of) proposition(-s) which provide grounds for drawing the conclusion.

Automatic recognition of relevant semantic units involves two tasks: (1) *segmentation* of a text into meaningful units; and (2) *annotation* of these units capturing (part of) their meaning. Many argumentation mining studies assume that the boundaries of the argument components have been previously detected by other means, thus they focus on the classification task (Stab and Gurevych, 2014; Eckle-Kohler et al., 2015). Other consider segmentation as a sub-task and perform both segmentation and classification (Levy et al., 2014; Rinott et al., 2015).

We define *argument structure recognition* to involve: (1) segmentation of a text into elementary argumentative units assuming that they correspond to elementary discourse units; (2) discourse relation detection between them; (3) classification of the identified relations; (4) classification of the identified argumentative units based on the classified discourse relations; and (5) argument completion – reconstruction of implicit units to achieve a complete argument structure, see also (Peldszus and Stede, 2013). In this study we evaluate state-of-the-art discourse parsers and machine learning models on automatic segmentation and discourse relation classification tasks, and then apply them to extract arguments from argumentative texts.

The rest of the paper is structured as follows. In Section 2, we discuss related work concerning argument structure recognition. Section 3 presents established discourse theories as theoretical and empirical framework for argument analysis and argumentation modelling. The connection to the existing ISO 24617-8 standard for discourse relation annotation is made. Section 4 discusses the performed experiments elaborating on the datasets, tools and outcomes. Section 5 summarizes the results and outlines the future research.

2 Related Work

Peldszus and Stede (2013) defined Argumentative Discourse Units (ADUs) as text segments corresponding to propositions that are argumentatively relevant and have their own argumentative function. ADUs reflect different ways to support a claim (Fig. 1), e.g. with the *basic* argument configuration consisting of a conclusion supported by exactly one premise, as in example (1) below. If there are multiple premises supporting a conclusion together, the structure is called *linked support* as in (2). Multiple premises which support the conclusion independently form a *multiple support* as in (3). *Serial* support links arguments to the conclusion where an argument contributes to further development of an already given argument (4). Peldszus and Stede (2013) consider the example shown in (5) to be a special form of support.

(1) [Books are better than TV.]$_1$ [Books enlighten the soul.]$_2$

(2) [Books are better than TV.]$_1$ [Books enlighten the soul.]$_2$ [They change your perspective on life]$_3$

(3) [Books are better than TV.]$_1$ [1. Books don't ruin your eyes like TV does.]$_2$ [2.Books allow your brain to imagine]$_3$ [3.Reading books can help you with spelling.]$_4$ [4.Reading books can help you write better.]$_5$

(4) [Gay marriage is wrong.]$_1$ [In fact, we would all become extinct,$_2$] [because without one man and one woman]$_3$ [there would be no reproduction.]$_4$

(5) [Personal pursuit is better than advancing the common good.]$_1$ [I need to think about me first, success and then think of others.]$_2$

Since not every text is argumentative and, therefore, subjected to an argumentative analysis, identification of its type can be considered as a preliminary step, and together with the topic context may provide valuable information for the argument component identification. Levy et al. (2014) introduced the notion of a *context-dependent claim* – a general concise statement that directly supports or contests a given topic. Rinott et al. (2015) detect *context-dependent evidence* – text segments that directly support a claim in the context of a given topic. Contextual information has served as an important source for argument component identification in

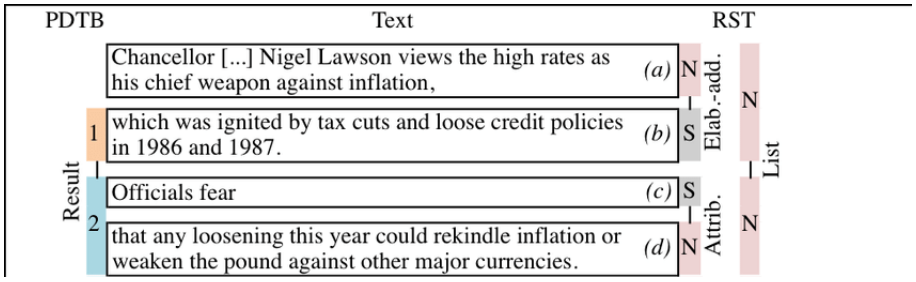


Figure 2: PDTB and RST-DT annotations for a *WSJ 1172* paragraph (Demberg et al., 2019), where 1 refers to Arg1 and 2 to Arg2 in PDTB; N stands for Nucleus and S for Satellite in RST; and (a-d) are RST-DT's Elementary Discourse Units.

Kuribayashi et al. (2018); Opitz and Frank (2019); Aker et al. (2017); Shnarch et al. (2018).

Mining arguments from diverse corpora based on topic can pose certain problems. A well-established topic is not always easy to determine or a text can cover several topics and the discussion can shift between them throughout the entire text. Lippi and Torroni (2015) proposed a method for *context-independent claim detection*. The approach relies on the assumption that argumentative sentences share the structure independently of the addressed topic. This technique was successfully applied for legal texts (Lippi et al., 2015), clinical trials (Mayer et al., 2018) and social media (Liga, 2019).

Cross-domain approach to the argumentation mining has been explored in a number of studies. Rosenthal and McKeown (2012) detect claims from two different data sets, LiveJournal and Wikipedia. Al Khatib et al. (2016) experimented with a wider range of text types and topics addressing politics, culture, religion, sport, economy, and health. Deep learning techniques were applied in cross-domain and multi-task learning scenarios (Eger et al., 2017; Daxenberger et al., 2017; Stab et al., 2018; Schulz et al., 2018; Morio and Fujita, 2019; Mensonides et al., 2019; Wambsganss et al., 2020).

Argument structure is often viewed through the prism of discourse theory and ADU components are defined based on discourse units which proves that argumentation and discourse characteristics, and these structures are closely related. Peldszus and Stede (2016) explored the mapping between discourse and argument(-ation) structures based on the Rhetorical Structure Theory (RST, Mann and Thompson (1988)) and those of Segmented Discourse Representation Theory (SDRT, Lascarides and Asher (2008)). Stede et al. (2016) assesses the role of discourse parsing features for argumentation structure prediction. Cabrio et al. (2013) and Hewett et al. (2019) translated the general simple argument structure into several discourse-based schemes to perform analysis and evaluation of natural language arguments, see also (Teufel et al., 1999; Palau and Moens, 2009; Petukhova et al., 2017). Eckle-Kohler et al. (2015) assessed the role of discourse markers for claims and evidence detection. In Hofmockel et al. (2017), the impact of the genre on different realizations of discourse relations is evaluated. Green (2018) applied the genre-based approach to scientific (e.g. biological/biomedical) texts.

3 Discourse Analysis

Discourse theory aims at explaining the coherence of a text. Its central notion is *coherence*, also called *rhetorical* or *discourse relation* - a semantic or pragmatic relation between two adjacent text spans. Even though text coherence and argumentation structure are not identical, discourse structure can reveal new unexplored properties of argumentation. Bridging from discourse to argumentation, Peldszus and Stede (2013) chooses the RST framework where all parts of a text are involved into a discourse structure and organized as a tree, with Elementary Discourse Units (EDUs) as leaves. An EDU is a minimal building block of a discourse tree which typically corresponds to a clause (Carlson and Marcu, 2001).[2] RST specifies how EDUs and larger units are connected, where some text spans are more important than the others, i.e. *nucleus* or multiple *nuclea* are the central part of a relation in the text supported by a *satellite*. The corresponding RST tagset contains 78 discourse relations which can be grouped into 16 classes sharing one type of rhetorical meaning.

Another influential discourse analysis frame-

[2]Other competing hypotheses take an EDU to be a prosodic unit, a dialogue turn, a sentence, an intentionally defined discourse segment (e.g. utterance) or the contextually indexed representation of information.

work is defined within Penn Discourse Tree Bank (PDTB, Prasad et al. (2005)). PDTB does not make strong assumptions about the overall structure of a text and does not suggest what kinds of high-level structures may be created from the annotated low-level relations and arguments. The PDTB analysis is focused on the discourse relation between two text segments called `Arg1` and `Arg2` which can be treated as EDUs. PDTB accounts for the lexical items that can signal discourse relations – discourse connectives. In the case of *explicit* connectives, `Arg2` is the argument to which the connective is syntactically bound, and `Arg1` is the other argument. In the case of relations between adjacent sentences, `Arg1` and `Arg2` reflect the linear order of the arguments, with `Arg1` before `Arg2`. PDTB does not constrain an EDU to be a single clause or single sentence, however, the framework follows a minimality principle requiring an argument to contain the minimal amount of information needed to interpret the relation successfully. The PDTB annotation scheme forms the basis of the ISO DR-Core (ISO 24617-8) discourse relations annotation standard (Bunt and Prasad, 2016).

Even though RST and PDTB annotation frameworks make different assumptions about the discourse structure and define different sets of relations, Demberg et al. (2019) suggest an automatic alignment of their relations and evaluates the mapping discrepancies. Figure 2 compares PDTB and the RST Treebank (RST-DT, Carlson et al. (2003)) annotations of the WSJ-1172 paragraph of Penn Tree Bank (PTB, Marcus et al. (1993)).

Discourse analysis within both annotation frameworks includes (1) segmentation of the text into EDUs; and (2) the recognition of discourse relations between these units. Discourse parsers typically perform both tasks. For example, Lin et al. (2010) designed a full parser to perform the PDTB annotations. The system first identifies discourse connectives, label the corresponding `Arg1` and `Arg2` spans and assign an `Explicit` relation. If no connective was identified, the system classifies the statement pair as having one of the other relation types, i.e. `Implicit, EntRel, AltLex, NoRel`.

Wang and Lan (2015) extended the parser with extractors for `Arg1`, `Arg2` and `Non-EntRel` relations. Qin et al. (2016) improved recognition of the implicit relations. Recent works explore deep learning techniques which use architectures for multi-task learning (Liu et al., 2016; Lan et al., 2017; Van Ngo et al., 2019) or adversarial neural networks (Qin et al., 2017; Huang and Li, 2019).

While many studies focus exclusively on the discourse relation recognition assuming that the text is already pre-segmented, others also consider discourse segmentation task. Early generation segmenters were rule-based systems (LeThanh et al., 2004; Tofiloski et al., 2009), whereas more recent approaches view this task as sequence labeling problem and use deep learning (Hernault et al., 2010; Bach et al., 2012; Wang et al., 2018). Multilingual discourse segmentation is addressed in (Braud et al., 2017; Muller et al., 2019; Desai et al., 2020).

The presented study concerns both segmentation and classification tasks assessing the performance of the state-of-the-art tools on the argumentative corpus. Design and results are reported in the next Section.

4 Experimental Design

We conducted the following experiments: (1) evaluating the quality of the existing full discourse parsers on EDUs segmentation and relation classification tasks; (2) two-stage discourse segmentation and relation annotation; (3) application and evaluation of the best performing model to identify and classify argument components in the argumentative corpus; and (4) completion of argument structure by reconstructing implicit claims. Figure 3 shows the experimental workflow.

4.1 Datasets

There are two corpora used in this study: *Penn Discourse Treebank 2.0* (PDTB 2.0, Prasad et al. (2008))[3] – a large scale corpus annotated with information related to discourse structure and discourse semantics, and *Dagstuhl15512 ArgQuality* (Wachsmuth et al., 2017) – a corpus of segmented arguments annotated with argument quality scores.

PDTB 2.0 consists of 2,159 articles from Wall Street Journal (WSJ) divided into 25 sections. In total, there are 40600 discourse unit pairs annotated with different relations. We provide a list of relations and their distribution in the Appendix.

[3]PDTB 2.0 is an extended version of the PDTB 1.0 corpus, where extensions concern annotations of implicit relations for the entire corpus, senses of all connectives and attribution of object type, scopal polarity and determinacy. Thus, for the purpose of this study, differences between PDTB 1.0 and PDTB 2.0 are not relevant.

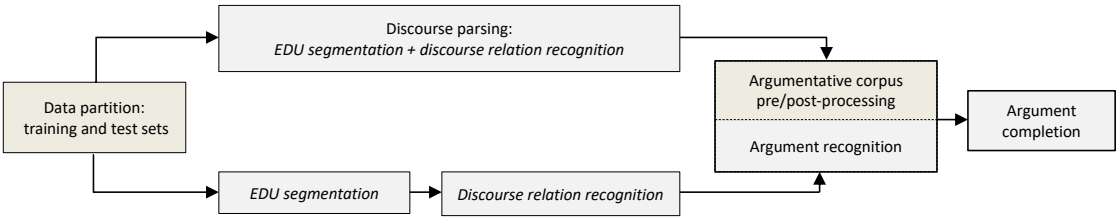


Figure 3: Experimental workflow for the argument structure recognition.

Dagstuhl15512 ArgQuality (Wachsmuth et al., 2017) is a collection of argumentative texts from the *UKPConvArgRank* dataset (Habernal and Gurevych, 2016) consisting of debate portal arguments *for* and *against* stances on 16 topics. The *UKPConvArgRank* dataset was developed to predict the convincingness of arguments, so that each argument pair is rated as more or less convincing. For *Dagstuhl15512 ArgQuality*, texts on each topic containing the five top- and five bottom-ranked arguments were selected and annotated across three core quality dimensions: argument cogency, argument effectiveness and argument reasonableness, and several sub-criteria (15 in total).

4.2 Discourse Analysis Tools Assessment

4.2.1 Discourse Parsing

One of the widely used tools for discourse processing is the PDTB parser developed by Lin et al. (2010). It is trained on sections 02-21 of Penn Discourse Tree Bank (PDTB 1.0, Prasad et al. (2005)) for text span identification and relation classification. For our purposes, spans for both `Arg1` and `Arg2` need to correspond exactly or partially to the PDTB 2.0 reference segments. Moreover, the relation between EDUs should be correctly classified. We evaluated the parser performance on the full PDTB 2.0 corpus. Table 1 summarizes parser performance in terms of F1 scores. The gold standard parsing and EDUs boundaries with error propagation setting ($GS + EP$) refers to a clean, per-component evaluation. In the automatic parsing and EDUs boundaries with error propagation scenario ($Auto + EP$), end-to-end automated parsing of the unseen data is performed. In the later setting, F1 scores of 38.18% and 20.64% were achieved for partial and exact match, respectively. A large portion of the misclassified cases belong to the `Non-Explicit` classes, as implicit discourse relations are more difficult to classify. The bottom part of Table 1 reports F1 scores obtained on the EDU span identification and on the joint segmentation and classification tasks on the entire PDTB 2.0 corpus.

Experimental setting	**F1 score** (%)
GS + EP (partial match)	46.80*
Auto + EP (partial match)	38.18*
GS + EP (exact match)	33.00*
Auto + EP (exact match)	20.64*
EDU span identification	22.61**
EDU span identification & relation recognition	21.20**

Table 1: Performance (F1 scores) of the PDTB parser developed by Lin et al. (2010) on various tasks. * evaluation performed on the section 23 of the PDTB 2.0 corpus; ** evaluation performed on the on full PDTB 2.0 corpus.

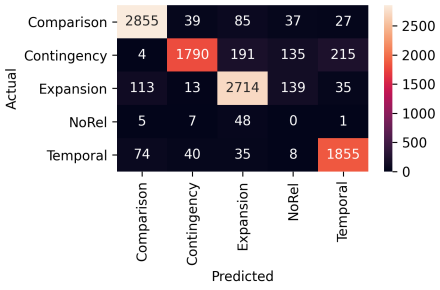


Figure 4: Confusion matrix for L1 relation classification with the PDTB parser.

Similarly to Hewett et al. (2019), we observed that the parser failed to identify many spans correctly. In case of the correct span identification, relation classification was reasonably accurate. Figure 4 shows the confusion matrix for the top-level (L1) relations between the correctly identified pairs of `Arg1` and `Arg2`. We concluded that the parser generally tends to assign a relation between the majority of EDU spans misclassifying `NoRel` instances.

4.2.2 Discourse Segmentation and Relation Recognition

As shown in the parser evaluation experiments, EDU segmentation is a crucial step in discourse analysis. Since the PDTB parser failed to show satisfactory segmentation performance, we tested state-of-the-art neural network model on the reference PDTB annotation, i.e. the BiLSTM-CRF

EDU segmentation			*PDTB Relation recognition*			
Statistics		**F1 score** (%)	**Statistics**			**Accuracy** (%)
			# Classes	Training set	Test set	
total segments	123780		2 classes	62172	4655	88.86
exact matches	13420 (10.84%)	68.55	5 classes	19145	4655	66.37
partial matches	56847 (45.92%)		10 classes	12070	4471	53.64

Table 2: Segmentation performance (F1 scores) with an overview of the exact and partial matches processed with NeuralEDUSeg (Wang et al., 2018) of the PDTB 2.0 corpus; and relation recognition accuracy for different classification scenarios applying the XLNet model (Yang et al., 2019) on PDTB 2.0 data.

based model NeuralEDUSeg developed by Wang et al. (2018). In this experiment, a unit is acknowledged to be correctly segmented if it partially or fully corresponds to one of the reference PDTB segments. Table 2 reports the number of exact and partial matches. Segmentation performance achieves 68.55% of F1 score. Our results show that NeuralEDUSeg significantly outperforms the PDTB parser (compare with Table 1). While the number of exact matches is still rather low (10.84%), we observed a relatively high number of identified partial matches (45.92%). The fact that most matches coincide with the reference segmentation only partially can be explained by the fact that NeuralEDUSeg is originally trained on the RST-DT corpus which follows different segmentation principles (consider Figure 2 again). Minimal RST-DT units tend to be shorter than those of the PDTB. For example, compare the NeuralEDUSeg [segment]$_1$ with the PDTB [segment]$_2$ illustrated in (6):

(6) a) [Woolworth said]$_1$ [Woolworth said it expects to expand usage of the MCI services as it adds about 6000 business locations over the next few years]$_2$
b) [The derivative markets remained active]$_1$ [The derivative markets remained active as one new issue was priced]$_2$

Deep learning models show promising results on discourse relation recognition task. Kim et al. (2020) demonstrated that the XLNet-large model of Yang et al. (2019) achieved the best results on implicit discourse relation recognition significantly outperforming BERT- (Nie et al., 2019) and ELMO-based (Bai et al., 2019) discourse relations models.

We performed a series of experiments on fine-tuning XLNet for the discourse relation recognition task. We first conducted a binary classification to establish whether is any relation between the identified units, i.e. the model discriminates between `Rel` class (includes any type of discourse relations) and `NoRel` comprising the `EntRel` and `NoRel` types. Secondly, we performed five-class top-level (L1) and ten-class fine-grained (L2) relations classification. The following five classes were used for the second experiment: *Expansion, Conjunction, Comparison, Contingency, Temporal, NoRel*. The ten-class experiment exploited the classes listed below: *Expansion.Conjunction, Expansion.Restatement, Expansion.Instantiation, Temporal.Synchrony, Temporal.Asynchronous, Contingency.Cause, Contingency.Condition, Comparison.Contrast, Comparison.Concession*. See Appendix for the class distribution. Classes with less than 500 training instances were excluded. The training set comprised sections 0-21 of the PDTB 2.0 corpus; sections 22-24 served as the test set. Since classes were not balanced in all classification settings, we performed *re-sampling* procedure: *up-sampling* of the under-represented `NoRel` class in binary classification by adding synthetic samples combining random EDUs from different textual units; and *down-sampling* the majority classes in the multi-class settings. The right part of Table 2 presents the final training and test data partitions for each classification scenario.

For the training and evaluation procedure, we fine-tuned each encoder model following the suggestions of Mosbach et al. (2021) and trained for 10 epochs using a learning rate of 0.00001 and a batch-size of eight. The results are summarized in Table 2, from which we can observe that accuracy drops with a higher number of classes to learn, from 88.85% for two classes to 53% for ten classes. We note that the results of our experiments differ from those reported by Kim et al. (2020) due to the differences in the approach and set of the classified relations. For instance, we were not focused on the distinction between implicit vs. explicit relation recognition. The goal was to assess how well the model predicts cases when a relation between two segments exists without focusing on how this relation is expressed. Moreover, we included `NoRel` instances into the classification, while they are typically discarded in other studies.

4.3 Discourse-based Analysis of an Argumentative Corpus

We applied the tested discourse analysis tools on *Dagstuhl15512 ArgQuality* corpus where we manually examined and corrected the model outputs. Respecting the PDTB minimality principle, we combined or split relevant text units depending on the amount of information required to interpret the relation between the segments correctly. We conduct a detailed error analysis and discuss some representative cases below.

We encountered many examples where a single unit does not contain substantial semantic information and has to be combined with the adjacent segment(-s) as illustrated in (7):[4]

(7) [A law] [requiring separate schools and public accommodations for homosexual people would violate] ["separate but equal"] → [A law requiring separate schools and public accommodations for homosexual people would violate "separate but equal"]

We considered modal constructions such as *I think, I believe, I am sure, Maybe, I highly doubt* as in (8), infinitive constructions (9), participle constructions (10) and relative clauses (11) as not forming an EDU on their own and therefore not having any discourse relation to the neighbouring EDU(-s). Relevant segments are merged.

(8) [I believe] [it should not be done] [just to discipline a child.] → [I believe [it should not be done just to discipline a child.]

(9) [Congress have no power] [to pass a legislation] [forcing religious institutions about marriage.] → [Congress have no power to pass a legislation forcing religious institutions about marriage.]

(10) [it doesn't break the Separation between Church and State] [ruled by the Supreme Court.] → [it does n't break the Separation between Church and State ruled by the Supreme Court.]

(11) [It would be hard for me to turn in the one] [I love.] → [It would be hard for me to turn in the one I love.]
[Yes, if the person] [I loved] → [Yes, if the person I loved]

We also encountered a few cases where the segment identified by the parser can be split into several EDUs as in (12):

(12) [So, many countries depends on scientists.most of employees in every country] [is Indians.], [and still be successful. Take myself for example;] → [So, many countries depends on scientists.] [most of employees in every country is Indians.]

EDU segmentation		*PDTB relation recognition*	
Match type	**F1 score** (%)	**# Classes**	**Accuracy** (%)
exact match	47.94	5 classes	60.22
partial match	79.83	10 classes	50.48

Table 3: Performance on EDU segmentation task applying NeuralEDUSeg model Wang et al. (2018) on the Dagstuhl corpus in terms of F1 scores (in %); and accuracy scores (in %) for 5- and 10 class discourse relation classification on the DagStuhl corpus with the fine-tuned XLNet-large model.

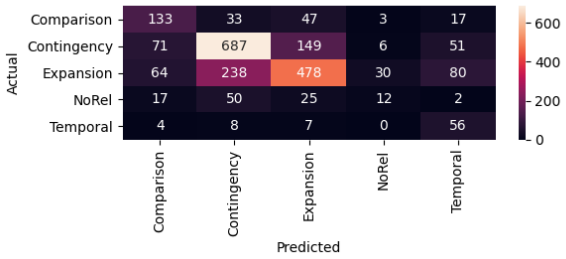


Figure 5: Confusion matrix for 5-class discourse relation classification task on Dagstuhl15512 ArgQuality.

Table 3 reports the performance of the NeuralEDUSeg model evaluated on the manually segmented argumentative Dagstuhl corpus using the reference segmentation.

As the next step, the identified EDUs were used to classify discourse relation between them applying XLNet (Yang et al., 2019). For this, pairs of adjacent segments were constructed and annotated with the PDTB discourse relations. We focused on ten classes mentioned above (the distribution is provided in the Appendix).

Wachsmuth et al. (2017) notes that some argument components, most often a claim, can be implicit. Consider an example in (13) below. An argument is not complete without the claim and cannot be used for further argumentation mining tasks. Therefore, we reconstructed a claim for every topic in the corpus, i.e. either 'for' or 'against' stance it may present. The reconstructed claim is a simple sentence which correspond to a single EDU. Subsequently, the reconstructed claims were used to created EDUs pairs for discourse relation classification.

(13) (a) The question is: who has the right to prohibit it? Government? Why would there be any pressing need at all for the state to outlaw pornography? Look at Europe- they're cool with pretty much everything. I don't see

[4]Here and in the following examples, a text span in the square brackets corresponds to an EDU obtained with a neural discourse segmenter; the manually corrected version is given after the arrow sign →.

any moral depravity in Europe, do you? (implicit claim: Pornography is not wrong.)
(b) Books will be always great whatever the new technological developments emerges, books has its fixed place in every humans heart. (implicit claim: Books are better than TV.)

EDUs pairs were built considering non-adjacent text units connected by a discourse relation. Most frequently, a claim may be connected to segments representing various types of evidence at different support levels as in (14):

(14) [Advancing the common good is better than personal pursuit.] [I think common good is better than personal pursuit]
[Advancing the common good is better than personal pursuit.] [When people help each other out its more likely that everything comes out great.]
[Advancing the common good is better than personal pursuit.] [Yes personal pursuit is important]

The resulting *Dagstuhl* corpus annotated with discourse relations contains the same number of 304 arguments as the original one which are segmented into 2,222 EDUs pairs. The XLNet-large model, initially trained and fine-tuned on the PDTB 2.0 dataset, was evaluated on Dagstuhl, see Table 3 for the performance overview. Figure 5 presents the confusion matrix for the 5-class relation classification task. We observed that many relations are correctly classified even in the absence of discourse connectives on which the model relies. Consider the following classification output:

(15) Creationism tries to sneak the supernatural as a scientific explanation. `Expansion.Restatement` This is called pseudo - science.

So a lousy father is better than none. `Comparison.Concession` (that is of course assuming that he is not abusive in any way)

Books enlighten the soul. `Expansion.Conjunction` Books don't destroy the morals of children.

and the big corporations like Dasani and Nestle would loose millions of dollars. `Contingency.Cause` It would hurt the economy severely .

Physical education does absolutely nothing for the children 's health and/or lifestyle . `Expansion.Instantiation` Let me describe my PE experience. Throughout my public education career , PE has been mandatory for each year.

I think common good is better than personal pursuit `Comparison.Contrast` Yes personal pursuit is important.

it wouldn't be so easily for you to become fat `Contingency.Condition` (of course you would also need to keep a balanced diet)

To summarize, the evaluated discourse processing tools showed a reasonable segmentation (F1 score ranging from 47.94% for exact match to 79.83% for partial match) and discourse relation recognition (accuracy ranging from 50.48% to 60.22%) performance on argumentative data. Thus, they can be applied in argument structure recognition and reconstruction tasks.

5 Conclusions and Future Work

The presented study reviewed discourse-based approaches to argumentative discourse analysis. We evaluated three widely used tools on argument segmentation and annotation tasks, namely, a rule-based PDTB full parser (Lin et al., 2010), a BiLSTM-CRF model for discourse units segmentation (Wang et al., 2018) and an XLNet based discourse relations classifier (Yang et al., 2019). Our experiments demonstrated that the PDTB parser achieved an F1 score of 22.61% on the span identification and 21.20% on the joint span identification and relation recognition tasks. This performance has been considered unsatisfactory for further use. Deep learning models, in contrast, showed significantly better performance: F1 scores ranging from 47.94% to 79.83% were achieved on the segmentation task, and accuracy of 60.22% and 50.48% for top-level and fine-grained discourse relation classification, respectively.

We successfully applied the best performing models to segment and annotate the argumentative corpus *Dagstuhl15512 ArgQuality* and conducted the detailed error analysis. The obtained argumentative discourse units were manually corrected and annotated with the fine-grained PDTB discourse relations. This corpus contains 2,222 annotated unit pairs and presents a valuable resource for further argumentation mining studies and will be released to the community.

The obtained results opened up many interesting prospects for future research. For example, various argumentation schemes can be reconstructed based on the proposed approach, and evaluated within numerous contexts and domains. Argument and argumentation quality can be assessed and robust reasoning engines designed.

Acknowledgments

The research reported in this paper was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 232722074 SFB 1102.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first workshop on argumentation mining*, pages 64–68.

Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, and Mina Ghobadi. 2017. What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 91–96.

Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1395–1404.

Kevin Ashley, Niels Pinkwart, Collin Lynch, and Vincent Aleven. 2007. Learning by diagramming supreme court oral arguments. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 271–275.

Ngo Xuan Bach, Nguyen Le Minh, and Akira Shimazu. 2012. A reranking model for discourse segmentation using subtree features. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL 12, page 160168, USA. Association for Computational Linguistics.

Hongxiao Bai, Hai Zhao, and Junhan Zhao. 2019. Memorizing all for implicit discourse relation recognition. *arXiv preprint arXiv:1908.11317*.

Xue Bai. 2011. Predicting consumer sentiments from online text. *Decision Support Systems*, 50(4):732–742.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017. Cross-lingual and cross-domain discourse segmentation of entire documents. *arXiv preprint arXiv:1704.04100*.

Harry Bunt and Rashmi Prasad. 2016. Iso dr-core (iso 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th joint ACL-ISO workshop on interoperable semantic annotation (ISA-12)*, pages 45–54.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 1–17. Springer.

Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. Do not feel the trolls. In *Proceedings of the 3rd International Workshop on Social Data on the Web*, Shanghai, China.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Vera Demberg, Merel Scholman, and Fatemeh Asr. 2019. How compatible are our discourse annotation frameworks? insights from mapping rst-dt and pdtb annotations. *Dialogue and Discourse*, 10:87–135.

Takshak Desai, Parag Pravin Dakle, and Dan Moldovan. 2020. Joint learning of syntactic features helps discourse segmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1073–1080.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104*.

Nancy L Green. 2018. Towards mining scientific discourse using argumentation schemes. *Argument & Computation*, 9(2):121–135.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A sequential model for discourse segmentation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 315–326. Springer.

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

Carolin Hofmockel, Anita Fetzer, and Robert M Maier. 2017. Discourse relations: Genre-specific degrees of overtness in argumentative and narrative discourse. *Argument & Computation*, 8(2):131–151.

Alexander Hogenboom, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska De Jong. 2010. Mining economic sentiment using argumentation structures. In *International Conference on Conceptual Modeling*, pages 200–209. Springer.

Hsin-Ping Huang and Junyi Jessy Li. 2019. Unsupervised adversarial domain adaptation for implicit discourse relation classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 686–695, Hong Kong, China. Association for Computational Linguistics.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414.

Tatsuki Kuribayashi, Paul Reisert, Naoya Inoue, and Kentaro Inui. 2018. Towards exploiting argumentative context for argumentative relation identification. In *Proceedings of the Annual Meeting of the Association for Natural Language Processing NLP*, pages 284–287.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308, Copenhagen, Denmark. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 2008. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Huong LeThanh, Geetha Abeysinghe, and Christian Huyck. 2004. Generating discourse structures for written texts. In *Proceedings of the 20th international conference on Computational Linguistics*, page 329. Association for Computational Linguistics.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Davide Liga. 2019. Argumentative evidences classification and argument scheme detection using tree kernels. In *Proceedings of the 6th Workshop on Argument Mining*, pages 92–97.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

Marco Lippi, Francesca Lagioia, Giuseppe Contissa, Giovanni Sartor, and Paolo Torroni. 2015. Claim detection in judgments of the eu court of justice. In *AI Approaches to the Complexity of Legal Systems*, pages 513–527. Springer.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. Technical report, University of Pennsylvania Department of Computer and Information Science Technical.

Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. 2018. Argument mining on clinical trials. In *Computational Models of Argument: Proceedings of COMMA 2018*, pages 137–148.

Jean-Christophe Mensonides, Sébastien Harispe, Jacky Montmain, and Véronique Thireau. 2019. Automatic detection and classification of argument components using multi-task deep neural network. In *3rd International Conference on Natural Language and Speech Processing*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230.

Gaku Morio and Katsuhide Fujita. 2019. Syntactic graph convolution in multi-task learning for identifying and classifying the argument component. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 271–278.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations (ICLR)*.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. Tony: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124. Association for Computational Linguistics.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 869–875. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):131.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 103–112, Berlin, Germany. Association for Computational Linguistics.

Volha Petukhova, Tobias Mayer, Andrei Malchanau, and Harry Bunt. 2017. Virtual debate coach design: assessing multimodal argumentation performance. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 41–50. ACM.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

Stephen Purpura, Claire Cardie, and Jesse Simons. 2008. Active learning for e-rulemaking: Public comment categorization. In *Proceedings of the 2008 International Conference on Digital Government Research*, pages 234–243. Digital Government Society of North America.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Shallow discourse parsing using convolutional neural network. In *Proceedings of the CoNLL-16 shared task*, pages 70–77.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. *arXiv preprint arXiv:1704.00217*.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on Empirical Methods in Natural Language Processing*.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *2012 IEEE sixth international conference on semantic computing*, pages 30–37. IEEE.

Matthew Rowe and Jonathan Butters. 2009. Assessing trust: contextual accountability. In *Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web*, Heraklion, Greece.

Olivia Sanchez Graillet and Philipp Cimiano. 2019. Argumentation schemes for clinical interventions. towards an evidence-aggregation system for medical recommendations. In *HEALTHINFO 2019. The Fourth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing*.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Manfred Stede, Stergos Afantenos, Andreas Peldzsus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1051–1058.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 77–80. Association for Computational Linguistics.

Linh Van Ngo, Khoat Than, Thien Huu Nguyen, et al. 2019. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4201–4207.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Marilyn A Walker, Pranav Anand, Rob Abbott, Jean E Fox Tree, Craig Martell, and Joseph King. 2012. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. 2020. Unlocking transfer learning in argumentation mining: A domain-independent modelling approach. In *15th International Conference on Wirtschaftsinformatik*.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 17–24.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. *arXiv preprint arXiv:1808.09147*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Appendix: Discourse relation distribution in PDTB 2.0

L1 top-level relations	L2 fine-grained relations	# Instances	
Expansion	***Expansion.Conjunction***	8763	15116
	Expansion.Restatement	3326	
	Expansion.Instantiation	1735	
	Expansion.List	627	
	Expansion.Alternative	531	
	Expansion	118	
	Expansion.Exception	16	
Comparison	***Comparison.Contrast***	5947	7958
	Comparison.Concession	1425	
	Comparison	553	
	Comparison.Pragmatic contrast	21	
	Comparison.Pragmatic concession	12	
Contingency	***Contingency.Cause***	6203	7710
	Contingency.Condition	1359	
	Contingency.Pragmatic cause	78	
	Contingency.Pragmatic condition	68	
	Contingency	2	
Temporal	***Temporal.Asynchronous***	2739	4352
	Temporal.Synchrony	1607	
	Temporal	6	
NoRel	***NoRel***		5464

Table 4: The PDTB top-level (L1) and fine-grained (L2) discourse relations and their distribution in PDTB 2.0 dataset. L2 relations in bold were used for 10-class classification with XLNet.

Converting Multilayer Glosses into Semantic and Pragmatic forms with GENLIS

Rodolfo Delmonte
Ca Foscari
University of Venice
Dept of Language Sciences
delmont@unive.it

Serena Trolvi
Ca Foscari
University of Venice
Dept. of Language Sciences
trolvi.serena@gmail.com

Francesco Stiffoni
Ca Foscari
University of Venice
Dept of Language Sciences
fstiffo@sgajo.com

Abstract

This paper presents work carried out to transform glosses of a fable in Italian Sign Language (LIS) into a text which is then read by a TTS synthesizer from an SSML modified version of the same text. Whereas many systems exist that generate sign language from a text, we decided to do the reverse operation and generate text from LIS. For that purpose we used a version of the fable *The Tortoise and the Hare*, signed and made available on Youtube by *ALBA cooperativa sociale*, which was annotated manually by second author for her master's thesis. In order to achieve our goal, we converted the multilayer glosses into linear Prolog terms to be fed to the generator. In the paper we focus on the main problems encountered in the transformation of the glosses into a semantically and pragmatically consistent representation. The main problems have been caused by the complexities of a text like a fable which requires coreference mechanisms and speech acts to be implemented in the representation which are often unexpressed and constitute implicit information.

1 Introduction

This paper presents work carried out for the automatic generation of written text in Italian language starting from glosses of fables in Italian Sign Language (LIS). The paper focuses on the semantic and pragmatic representation that has been created by the system GENLIS that feeds the generator. Whereas many systems exist that generate sign language from a text (Lombardo et al., 2011; Morrissey and Way, 2013; Wu et al., 2001; Sáfár and Marshall, 2001), we decided to do the reverse operation and generate text from LIS. A number of systems exist for American Sign Language that have attempted the same operation but only on a simple sentential basis and starting from visual recognition ((López-Ludeña et al., 2013; Dreuw et al., 2011; Elliott et al., 2000; Efthimiou et al., 2010)). The possibility to produce glosses automatically from video capture and image recognition (but see(Dorner and Hagen, 1994)) is not available for LIS, so we chose not to tackle the visual recognition phase and to start directly on the output, i.e. glosses[1]. Glosses for LIS are partly domain dependent in the sense that annotating sentences is a different task from annotating a dialogue, and this in turn is different from annotating a story or a fable. Among the many types of text that we could work on we chose the most difficult one: a fable, which is a mixture of narrative text and dialogues. For that purpose we used a version of the fable *The Tortoise and the Hare*, signed and kindly made available by *ALBA cooperativa sociale*. The signed story was annotated manually into glosses by second author - who is a LIS translator - for her Master's thesis (see also (Trolvi and Delmonte, 2020)). The fable has two main characters - and other secondary characters - with totally different personalities which may interact in dialogues, or may be simply narrated thus producing an overall complex textual structure.

2 Semantic and Pragmatic Representations from Glosses

As will be explained below, main problems have been caused by the complexities of a text like a fable - which is partly a dialogue and partly narration - and requires coreference mechanisms and speech acts to be implemented in order to convert glosses into a semantically and pragmatically consistent representations. The final text is organized into Discourse Units (hence DUs) or turns where each one may contain one or more sentences, and is associated with a unique turn identifier and a unique

[1] Transcription into glosses is a topic of research in itself because it may be done in different manners (Slobin et al., 2001; Hoiting and Slobin, 2002)

Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pages 54–64

speaker. Eventually we came up with 30 DUs, 54 sentences and 91 propositions. The full project is presented on a website *https://genlis.vercel.app/*. The website contains full representations for the all the DUs of the fable, showing the conversion process step by step. Every DU starts by the video clip of the actor performing the LIS narration of the current DU; this is followed by the multilayered annotation[2] which is then turned into the 9 slot prolog consistent vector-like term. The transcribed vector is then enriched by semantic information and then by pragmatic information. The final step is the Italian sentences produced by the generator[3], which are then spoken aloud by the speech synthesizer on any Mac or PC. The final part of *GENLIS* addresses the speech synthesizer with a set of prosodic markers to induce correct pauses, voice volume, intonational movements[4]. For lack of space we cannot comment on this part of the system: we can only say that we are using SSML on available speech synthesizers that accept it, to produce an expressive and semantically correct recital of the story. State-of-the-art generation systems work mostly on the basis of a machine learning approach (Stein et al., 2012), (Zhao et al., 2000), which crucially requires an adequate amount of training data to feed the model. In our case training data are not available[5] also because glosses for LIS are partly domain dependent as said above. In our case, we decided to generate text from a LIS version of the fable **The Tortoise and the Hare** which has two main characters with totally different personalities. As will be clear from the sections below, we decided to follow a traditional approach which apart from the starting phase - content determination made available by the glosses - continues with text structuring, sentence aggregation, lexicalisation, referring expression generation, and linguistic realisation. These phases could also be understood as the sequence of processes of ATLAS project (Lesmo et al., 2011), which however had the opposite task – thus a reversed input-output, i.e. generating LIS from Italian texts.

Generating text from manual multilayer glosses is different from traditional NLG (Natural Language Generation). Generation from LIS glosses does not follow from well structured data-sets or knowledge basis, nor is there a plan in order to build logically well-formed representations (Gatt and Krahmer, 2017). Glosses are mainly sequences of lemmata with some indication of plural number, negation, quantifiers with agreed features, numbers, personal pronouns. But then verbal, nominal and adjectival expressions are just lemmata, auxiliaries are missing and the same applies to copulative verb "to be" (Chesi et al., 2008). There are eight layers which specify type of speech act, presence of spatio-temporal location adverbs, role of current turn taker. They need to be collapsed and accounted for in the conversion phase in order to organize predicate-argument structures with all available information and converge towards a discourse level semantic and pragmatic representation.

GENLIS is written in the logic programming language Prolog (Gal et al., 1991; Mellish et al., 2006; Reiter, 2010), which makes available DCG (Definite Clause Grammar) rules together with Difference Lists to support text generation. The sequence of processes carried out by the system are represented in Figure 2 below.

Semantic forms are composed by main predicate, propositional attributes (such as e.g. mood, negation, verbal tense), arguments and adjuncts. Furthermore, each argument has its own internal structure. Semantic forms constitute the string that is eventually fed as input to the generator and then processed, in order to generate Italian sentences. We will describe below both the process of conversion of glosses into semantic forms and the structure of semantic forms. We will skip the first step in the whole process, which is producing the glosses and is done manually. As described in detail in another paper (Trolvi and Delmonte, 2020), manual glosses may contain arbitrarily many layers but they have the goal to interpret the signs in a shared manner. They are basically multi-layer text annotations written in tables, which can be done using one of the many software

[2]Manual annotation of simpler texts - either narrative or conversational - is not a highly time-consuming activity and can be carried out by an expert in a relatively short time.

[3]We are not aware of the existence of many generators for Italian (Lesmo et al., 2011) apart from the ones built by some of our collaborators (see (Delmonte and Bianchi, 1998; Delmonte and Pianta, 2008)) who were also partly the authors of a smaller version of the current one. The generator is now a general tool to generate most Italian sentence structures, and has been used in a number of other applications, like question-answering from a Discourse Model (see (Delmonte, 2000)).

[4]Intensive work on speech synthesis has been done in the past and also currently (see (Delmonte, 2016))

[5]Parallel corpora LIS-Italian text are available in a small number: besides ATLAS project (Lesmo et al., 2011), there is (Chesi et al., 2008) and (Barberis et al., 2011), none of which, however, will suit the genre requirements of the fable.

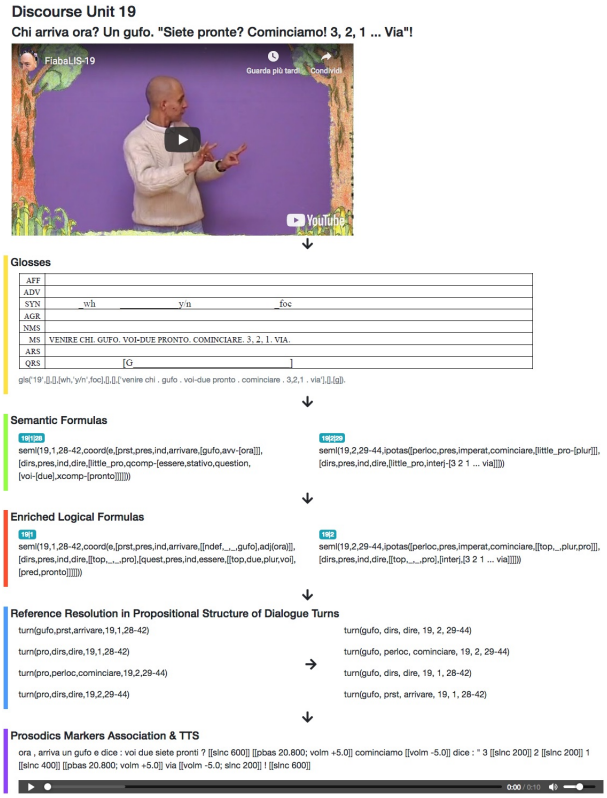


Figure 1: Snapshot of Discourse Unit 19 as presented by the website https://genlis.vercel.app dedicated to the generator.

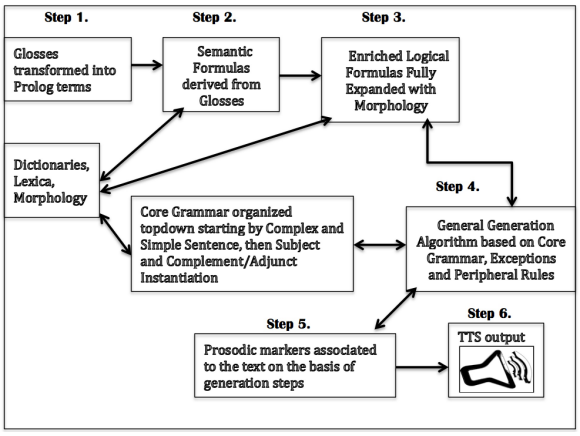


Figure 2: Flowchart of the GENLIS system decomposed into 6 steps.

available for the task - at first we used ELAN[6], but then we produced our own schemes to suit the requirements of the generator. In fact, in order for the multi-layer glosses to be analysed by the generator, it has been necessary to transform them into a 9 slot Prolog term. Thus, each annotation tier has been inserted in slots in a term, as follows:

[6]It can be downloaded here https://archive.mpi.nl/tla/elan - visited on April 2021

gls(DUInd,Aff,Adv,Syn,Agr,Nms,Ms,Ars,Qrs)

where the functor gls is an abbreviation for gloss, and contains a sequence of 9 slots explained here: DUind is the Discourse Unit index; the slots Aff, Adv, Syn contain annotated information about affective, adverbial and syntactic Non-Manual Signs; Agr identifies location and agreement of signs; Nms and Ms contain Non-Manual Signs and Manual Signs respectively and are expressed in a tokenized sequence between apostrophes ' ' as atomic objects; Ars and Qrs identify the occurrence of Action Role Shift and Quotation Role Shift[7].

2.1 The Conversion of Glosses into Syntactic/Semantic Lexical Forms

When creating conversion rules, we avoided indicating specific features that would make forms difficult to read and understand. More precisely, we did not indicate tense, mood and diathesis of verbs, number and genre of nouns and semantic role of oblique arguments for the generation of prepositions. We decided conventionally to generate sentences with active diathesis, in past tense and indicative mood. However, there are several factors to take into consideration: direct speech and questions, for example, are always expressed in present indicative. Furthermore, morphological features of nouns are always singular, unless otherwise indicated in glosses, and gender is derived from lexical gender. Past verb tense is derived on the basis of aspect of lexical verb; in particular, state and action verbs are expressed in *imperfetto* (a tense existing in Romance languages but not in English) tense, and the other verbs in past tense (*passato remoto* in Italian). Every fully expressed proposition has a verb that needs semantic and morphological features. While Person, Number and Gender may be inherited from the Subject, Tense and Mood are

[7]Role Shift is one of the main topics of the paper published on annotation of the fable (see (Trolvi and Delmonte, 2020)). It is a particular narrative strategy by which the signer adopts the perspective of another referent. Role Shift can be used to report a speech or thought of a referent or to reproduce his or her actions, thus it can be divided into two varieties. The terminology for both phenomena is not consistent throughout the literature. In our work, we adopted the terminology used by (Herrmann and Pendzich, 2018), namely "quotation role shift" (QRS) and "action role shift" (ARS). Hence, QRS is the type of RS by which the signer reports words or thoughts of other referents. ARS allows the iconic reproduction of actions, mannerisms and emotional states, including facial expressions and non linguistic gestures. It involved the use of the upper parts of the body (e.g. torso, head, eye gaze).

semantically and pragmatically determined. We have used lexical properties and discourse related (pragmatic) properties to assign Tense and Mood together with general consideration defined on the basis of narratological criteria. A fable or children story may be expressed using Indicative Present or Past tense (or *passato remoto*), however contextual conditions may impose constraints that require other Mood and Tense to be assigned. We may need to use Future tense, Imperative mood, Past tense (or *passato remoto*) rather than Present tense. A first subdivision of Mood-Tense assignment depending on Speech Act is shown below, a second subdivision follows according to Lexical Aspectual properties.

- Presentative constructions
  Perlocutive utterances
  Question + Exclamation
  Illocutive constructions
  Direct Speech constructions
  Statements

We distinguish Perlocutive from Illocutive verbs on the basis of the pragmatic nature of the action expressed: instructions on how to carry out a task are tagged Perlocutive and are enacted with Imperative mood. Illocutive expressions are tagged when the utterance expresses a decision or a wish to come true and are placed in the future Tense. Then, as a general rule, Activities are realized with Indicative Imperfetto, while Achievement use Past-tense (passato-remoto). The remaining cases are all realised with Indicative Present.
Semantic forms are structured as Prolog terms. Consistent with First-Order Logic (FOL), each term represents the content of a semantic proposition and is preceded by the functor PROP. PROP is the abbreviation for *proposition* and contains a fixed number of slots that mark semantic and pragmatic components included in glosses. More precisely, in first slot we may find pragmatic components like interjections - for expressing surprise or other affective and emotional aspects-, intrasentential elements like discourse markers and adverbs with scope on the verb or on the entire sentence. Let us now focus on the arguments structure. With the exception of SUBJect and OBJect, arguments are introduced by a functional marker that we derive from LFG theory (Bresnan, 2002), such as OBL for oblique arguments, FCOMP for sentential complement, VCOMP for verb complement and XCOMP for predicative complement. Moreover, argumental heads may contain modifiers, which are introduced by the marker MOD, or specifiers, which are usually included in brackets. If the argument is an expression of the affirmative or the negative polarity, the marker becomes the only term of the argument list. Moreover, direct speech is usually deprived of any introductory verb, which needs to be generated in Italian instead and may assume different meanings depending on context, as we will see in the next sections. Conversion rules from manual glosses are shown below:

- Identify elements that modify the main predicate, adverbs or discourse markers
- Insert the first verb you find
- Retrieve lexical verb aspect and create mood/time matrix
- The verb may be preceded by a location, which may be marked by a specific deictic term on the basis of type
- Speech act may vary
  PRESENTATION = WHO?
  DIRSPEECH for direct speech
  QUESTION if the sentence is a question
- Insert arguments into a list

Generate nominal expressions: The subject in first slot may be unexpressed. If so, it is marked with little-pro[8]: morphological features are retrieved from the subject of previous sentences. In case of direct speech, arguments may be interjections or statements/negations. The object may be a complement sentence marked FCOMP, an interrogative complement sentence marked QCOMP or an infinitive sentence marked VCOMP. Oblique arguments or adjuncts are marked OBL and in their first position they may contain either a preposition, if expressed overtly in manual glosses, or a semantic marker, and the lexical head in their second position. Nouns may have specifiers, such as *quale*/which in gara-[quale] (translated as race-[which]) and modifiers, which are marked MOD. Adverbs such as locative deictic adverbs are

[8]This label is derived from the Chomskyan linguistic theory that assumes the existence of an empty pronominal in pro-drop languages like Italian carrying morphological features derived from the main verb in sentences where the subject has been dropped, a choice which can be freely made in Italian and is based on discourse properties.

marked AVV. Gerundives are marked AVV too and contain the corresponding verb in infinitive form. PROPositions may be coordinated (COORD) or appear in sequence without markers (IPOTAS). These tags are inserted first, before the PROP tag. Examples are visible always in Figure 1 above.

All nominal expressions - both SUBJect and OBJect and OBLiques - can be modified by simple modifiers, multiple modifiers, and relative clauses. All of them are structurally attached to the nominal head because they are semantically and morphologically dependent on the head. In fact, adjectivals require feature agreement, which needs to be restricted before generation in order to prevent failures. As to relative clauses, their internal arguments may require the same type of information, in particular, in case the argument controlled by the relative pronoun - which may be unexpressed - is the SUBJect. Relative clauses may also be governed by an adjunct relation, but this is not the case in our story. In order to realise the appropriate word forms, the morphological features of the nominal head governing the relative clause are passed to the clause level as BINDER bundle of features, which may be used by the Verb Complex and realized as SUJBect or OBJect features.

Generate Verbal Complex and Complementation: The verb complex receives semantic and morphological information from the subject if present, be it a nominal or pronominal head, or simply an empty subject which however may have morphological features, person, number and possibly gender. Choosing the correct verbal complement structure may be dependent on subject semantic categories, which are also passed to the verbal complex. Semantic features are checked by matching subcategorisation information stored in the lexicon for each possible structural outcome. For instance, a verb like *dire*/say has a multiple entry in our computational lexicon with four different complement:

- vcomp = INFINITIVAL
  ogg = DIRECT-OBJECT
  ogg2 = INDIRECT-OBJECT(dative)+f/fcomp
  = SENTENTIAL-OBJECT
  f/fcomp = SENTENTIAL-OBJECT

They are all characterised by the same general lexical category, TRANSitive, and the same conceptual and semantic category, *report-dir* - that is a reporting verb that can be used also for direct speech introduction. This also applies to other verbs that may undergo intrasitivisation like mangiare/eat, but also to verbs with different complement structures but identical categorisations, like considerare/regard and dipingere/paint. In particular, *considerare*/consider has an open complements like NCOMP (a nominal predicative complement) or XCOMP (a label for generic open complements including infinitivals). All open complements require morphological features to match, and this will allow for complement structures to impose agreement for those features. This can be different for other verbs where lexical category may vary, as is the case for accennare/hint that may change from intransitive to transitive; or for a verb like apparire/appear that may change from copulative to unaccusative. Our lexicon is organised around a limited number of entries, around 1000 for most frequent lexical entries according to frequency dictionaries[9], and another extended set of manually annotated entries, around 9000, for the remaining less frequent but always non rare entries, which have a different feature and argument organization. Aspectual categories are very important - as said above - in the choice of verbal morphology regarding Tense and Mood; while semantic and conceptual class may also be relevant in case a sentential complement is present, as will be clarified below. Another important feature of verbal complex is the requirements it poses on auxiliary choice and precise morphological information as to the Tense and Mood to be realised. In particular, simple vs. composite verbal complex may be realised, which in turn require specification of the appropriate auxiliary verb: *essere* for passive, reflexive, inherent reflexive and unaccusative classes, *avere* for active transitive and intransitive classes. Morphological information from the SUBJect is also required in case of auxiliary essere in order to generate the appropriate past participle. The same is required from the OBJect in case of pronominalization processes of the nominal head into a clitic pronoun, which however requires decisions that can only be made by a full-fledged pronoun resolution system - which is not implemented in the generator. As

[9]The list is derived from previous work on Italian Frequency Dictionaries, see (Delmonte et al., 1996)

to Person, this may be available in case the SUBJect is lexically expressed. Empty pronouns on the contrary do not realise Person feature, which is by default set to 3rd. Special cases are constituted by Imperative mood and Direct Speech. Imperative mood requires 2nd person to be realised if the command or instruction is addressed directly to the interlocutor. But there are commands in the fable addressed by the owl to both competitors, the hare and the tortoise, to start the race. In this second case, 2nd person plural is required. However, 1st person plural is also acceptable. Introducing 2nd person is not an easy task and we haven't been able yet to find a linguistically motivated trigger to do it. The verb is checked for agreement with SUBJect morphological features. This may cause failures in the generation step, until the appropriate verb form is produced.

Complements and adjuncts are selected according to the shape of the semantic form: nominal and sentential complements are made up of a four or five slots list, while an oblique may be constituted by a list containing five or six slots; a simple modifier has only two or three slots. Finally adverbials or interjections consist of one or two slots but contain a special label as unique identifiers. Sentential complements may be simple sentences preceded by a complementizer, which is locally generated; or they may be direct questions. In this second case, a question mark is added at the end. The two complement types are marked by a special label identifier FCOMP and QCOMP. A special case may be constituted by WH- questions as sentential complements, requiring a local WH- expression to be generated before the verb also in case it is an adjunct - i.e. when, how, where. These pronouns would be positioned after the verb in the logical form built from semantic forms. So they need to be raised, i.e. removed from the complement structure and generated in the appropriate position.

Semantic Conversion Rules for Peripheral structural Representations: Peripheral structures are those special stylistically marked structures, like Subject Locative Inversion with presentative structures, and complements realised as clitics, which need to be positioned before the verb. In both cases we implemented the rules to act at the end of the generation process. A SUBJect-Locative is used in the first sentence of the fable, when the hare is presented and appears on the scene as living in the woods. This is a typical introductory sentence for many fables or children stories and has all the required linguistic features: the protagonist is unknown and is realised as an indefinite nominal structure; the verb is unaccusative or intransitive. In this case *vivere*/live is used intransitively; the sentence is completed by presence of a Locative adjunct, *nel bosco/in the wood*. The main linguistic elements are all generated in their base structure, they are identified and displaced in order to produce a presentation structure where the Locative comes in first position followed by the verb complex and then comes the subject nominal and finally the rest of the sentence, which in this case is an apposition. The second rule of Inversion regards the well known Subject/Object Inversion in Direct Speech utterance where what is being said is positioned before the governing communication verb. For example, in the utterance DU.11 *Si', si', qui, rispose la lepre/Yes, yes, here, replied the hare* the generated sentence has the so-called deep order, Subj GovVerb Obj(the spoken utterance). The peripheral rule has the task to invert Obj and Subj and obtain the more naturally pronounced utterance, where the most important part (what is being said) comes at the beginning.

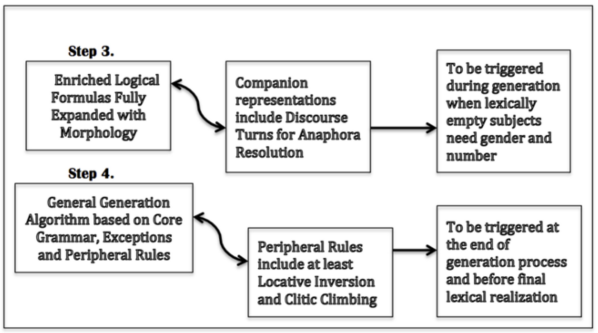


Figure 3: Peripheral Rules activated during Conversion and Generation.

The second case of peripheral rule is the one involving the generation of a clitic pronoun *ci* for a locative or a dative repeated in the same complement structure, and the governing verb *partecipare*/participate. The clitic is generated after the verb and then it is scrambled before it. Structures that require special rules to be implemented include so-called Open Complements and Open Adjuncts. Open Complements are predicative complements of copulative verbs, as in *siete pronti*/are you ready; Open Adjuncts are state adjectives like *tranquillo*/quiet, which require gender/number agree-

ment with the SUBJect as in *la tartaruga guardava tranquilla*/the tortoise was watching quiet. Both cases require SUBJect morphological features to be visible in the Complement/Adjunct section of the generator in order to select or restrict the appropriate word form.

3 Special Rules required by Implicit Elements

There is a number of rules that need to be organized mainly inside the conversion portion of the system. These rules regard a number of specific features that are missing in the LIS glosses and in the sign language as a whole. They concern *Definiteness Assignment*, that is the need to add an article in Italian sentences in front of a nominal expression, which could also be zero article. Then there is the need to vary the direct speech introductory verb which is otherwise always reported as DIRE/say. Eventually there is the need to map Tense, Mood and Person/Gender/Number onto all verb complexes.

3.1 The algorithm for Definiteness Assignment

In order for the generation to work properly, the feature *definite*, *indefinite* or *zero* must be decided automatically and inserted in the list of features associated to each nominal expression, be it the primary head as with subjects and object, be it secondary with obliques where the noun phrase is governed by a preposition. The list of features includes morphological, semantic and informational features as follows:

[Def,Spec,Num,Head]

Def contains the information about definiteness if the head is a noun, otherwise it is substituted by TOP in case the head is a pronoun. Spec contains information on quantification and any linguistic element that may be expressed by a quantifier. Num is associated to the morphological feature of Number. The Algorithm for Definiteness Assignment (ADA) is based on two parameters: the type of constituent and the semantics associated with the noun. The semantics is taken from a set of different sources due to their dimensions, which are insufficient to cover all nominal expression of the fable. We have been using the lexical-semantic database ItalWordNet(see footnote below), and the list of semantic general categories annotated therein. In the algorithm the main call is known-def, which is used to memorize the type of definiteness associated to a nominal head. When a noun is met for the first time it is asserted as NDEF i.e. indefinite, unless it belongs to a set of exceptions and special semantic classes. The choice of zero definiteness applies to nominal expressions characterized by an abstract feature, which in ItalWordNet[10] is represented by MNT (= mental) and EXPR (= expressive) tags. It also applies to words indicating location tagged by PART (= part) and PLAC (= place). Another interesting class is constituted by words belonging to Body-Part like *orecchio/ear*, which are tagged as definite and characterized by features PART, LIV (= living) and FNCT (= function); the same applies to nouns belonging to TIME semantic class, like days, months, but also *appuntamento/date* whenever they are included in a nominal constituent. The list of these nominals in our fable includes the following words:

appuntamento, vergogna, orecchio, giro, sinistra, destra, primo, tono/date, shame, ear, turn, left, right, first, tone

In addition, glosses' expressions like *referente-N* where there is a number varying from 1 to 2, are treated as pronouns. Frozen expressions like *3 2 1 ... via/3,2,1...go* are marked with definiteness zero. The number belonging to the class of ordinals is tagged with zero definiteness only in case they are included in an oblique governed by *arrivare/come*. Of course, all adverbial like expressions and interjections are not considered and do not receive a list of morphological and semantic tags as said above.

3.2 The Algorithm for Narrative direct speech speaking verb type

Discourse level processing is the most complex part of the algorithm, because it is responsible for overall discourse coherence and cohesion. In the glosses, direct speech is introduced always by the same verb *dire/say*. It may also be deprived of any introductory verb, which in our case needs to take into account the semantic content of the current utterance. In addition, depending on current discourse turn speaker, this verb may assume different meanings, which are strictly discourse

[10] https://www.cnr.it/it/banche-dati-istituti/banca-dati/442/italwordnet-iwn

related. So either *dire/say* is substituted by a contextually determined verb or a verb is introduced which was not present. These verbs belong to the Answering semantic type and are: *rispondere/reply* or *replicare/reply*, in case the speaker is answering a question from previous discourse turn. Otherwise the predicate may belong to the Asking type, *chiedere/ask* or *domandare/ask*, in case the current turn is made of a question; eventually it may also be *dire/say* in case the previous turn was a yes/no question, or the current turn is a statement. Finally with exclamations it may be *esclamare/exclamate*.
The algorithm is part of the convert file, the conversion algorithm that starting from glosses organizes them into semantic forms. It is activated after all conversions have been already made. The call is intended to modify the current predicate in case it is needed by the context. This is done checking semantic forms. Each turn is a vector representation, with current topic speaker, current speech act associated to current utterance, and a main predicate. The main predicate is headed by a Discourse Unit index, a Sentence index and a proposition index, like this: Head-Spac-Pred-Du-Sn-N. These representations are asserted into memory in a Prolog database and may be extracted easily.
The conversion algorithm receives Semantic Forms and checks to verify whether the current verb is *dire/say*. It also contains the current governing predicate, the arguments of current predicate in the Body variable, the Arguments of the first sentential complement (if any) of the Body variable in the variable Args, and finally the variable NewBody that will contain the modified version of the arguments. The first call to modify the predicate checks to see what is the speech act of the first proposition chosen. In this case the Predicate is substituted by a predicate of the Asking type, *chiedere-domandare/ask*. The second call is the most important one and is accompanied by a check of the previous turn. The call to verify previous turns is used to look into the database of turns. The search is interrupted in case the current utterance contains a question as one of its sentential complements. Then the second call searches the turns database. At first it extracts the previous turn and then it checks to see whether the current topic is different from the one asserted in the previous turn; finally it checks whether the speech act is a question. In this case the main predicate is modified into one of the Asking type. Eventually, the output of the generator is semantically coherent and pragmatically correct but it is fairly different from the one we created to stylistically suit a typical fable and interpreting the signer. Consider for instance the output of the generator for *DU 19: Ora arriva un gufo e dice : voi due siete pronti ? 3 2 1 via./Now comes an owl and says: you two are ready ? 3 2 1 go*. Compared to the utterance manually built corresponding to stylistically suit a typical fable story: *Chi viene ora? Un gufo. Siete pronte? Cominciamo! 3 2 1 ... via!/Who is coming now? An owl. Are you ready/fem/plur? Let's start! 3 2 1 ... Go!* This is shown in the figure below which is an excerpt from the website:

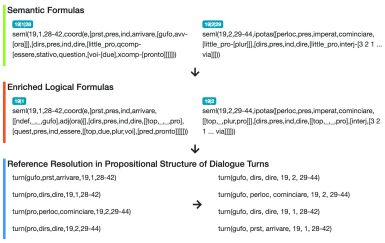


Figure 4: Excerpt of DU n.19 showing only the semantic and pragmatic conversion steps.

4 Evaluation and Discussion

Evaluation can be done manually or automatically (Belz and Reiter, 2006; Novikova et al., 2017). In order to do it automatically one would need a corpus of fables to be used for training which we currently don't have available. One should also take into account the need to measure how well glosses for LIS have been created and have been used by the system to produce a naturally sounding Italian text which resembles a fable. Also this evaluation is difficult to make for the same reason. We turned to human evaluation for lack of a better opportunity now left for the future. In order to evaluate the output of the generator we wrote manually a version of the story which was more adherent to what is expected from children fables and is attested in online versions of this fable. At the same time, the made up version had to respect as faithfully as possible the signed version produced by the signer in the video. The result is a story which is pleasant to listen to by children and adults, as we tested in a primary school classroom for an experiment. Now comes the evaluation of the generated story that we are able to produce by a comparison with the manually

created story - that we make available in full in the supplementary materials. The comparison was done at the beginning in order to produce the peripheral rules presented in the section above. What has been left unchanged is discussed here below. We decided to grade each Discourse Unit or SubUnit by a four levels graded scale: 1 = No Difference, 2 = Slight Differences, 3 = Noticeable differences, 4 = Very different.

1 = No Difference
No Discourse Unit or SubUnit is totally identical

2 = Slight Differences
a) Definiteness Assignment in DU. 1 un $\neq$ il, una $\neq$ la
P. In un bosco viveva una lepre, una lepre altezzosa.
G. Nel bosco viveva una lepre la lepre altezzosa.
We discuss this point using the first Discourse Unit of the story which we show here in the P(roposed) form and the G(enerated) form. The rule we created regards certain words as generic nominals which do not need to be individuated in the world and are assigned definiteness as they appear. This is the case of *bosco/wood*. The case of *una lepre/a hare* is different: at first appearance the nominal *lepre* is correctly assigned an indefinite article (una/a); as to second appearance our system computes HARE as already known in the world and assigns definiteness (la/the). But in this case the syntactic function of apposition reverts the semantics, because the apposition is just a means of characterizing the entity with additional attributes or properties. However this is difficult to realize in the generator.
DU1, DU7

b) Different Mood/Tense Present vs. Past Tense in DU. 2 avvicinò / avvicina
P. La lepre le si avvicinò ...
G. La lepre si avvicina ...
The rule for Mood/Tense assignment is sensitive to aspectual classes and speech act and we don't have the possibility to revert Present Tense to Past Tense in this case
DU.2, DU.3, DU4, DU5, DU8, DU15, DU18.2, DU22, DU23.1, DU23.2, DU25, DU26, DU27.2, DU28, DU29, DU30

c) Dative Ethic in DU. 2 le??
P. La lepre le si avvicinò ...
G. La lepre si avvicina ...
Presence of a Dative Ethic in Italian is optional and does not contribute to modify the semantics. We don't know of a linguistically motivated rule which could be used to insert it and make the sentence sound more natural
DU2

d) Use of a different direct speech communicative verb domandò/chiede
P. La tartaruga perplessa domandò...
R. La tartaruga chiede con aria perplessa...
DU3, DU4, DU5, DU9, DU10, DU11, DU12, DU17, DU25

e) Use of a different wh- word Che $\neq$ quale
DU6,

f) Use of a different locative adverbial lì $\neq$ qua in fondo $\neq$ là
DU7.1, DU7.2, DU27.1

g) Use of a different but fully synonymous verb from the one signed and inserted in the glosses
DU7.3, DU16, DU23.2, DU27.1

h) Use of a different exclamation interjection from the one signed
DU12

i) Presence of additional material in the generated story which was however present in the glosses and has been erased by the manually created story because redundant
DU13, DU14, DU27.3, DU29, DU30

l) Deletion of governing communicative verb in the generated story
DU10.2

3 = Noticeable differences
a) Omission of predicates present in the glosses
DU16.2

b) Omission of linguistic material like personal pronouns needed to reinforce the assertion
DU16.2, DU18.2

c) Mistaken gender associated to subject noun phrase or predicative open complement in copulative structures I-masc-plural due/Le-fem-plural

due
P. Le due si affiancarono.
G. I due si affiancarono.
DU18.1

d) Insertion of additional linguistic material in the manual story which was not present in the glosses DU1.2, DU7.1, DU18.2, DU18.3, DU19, DU20, DU21, DU23.1, DU27.2, DU29

e) Presence of linguistic material which is semantically almost synonymous but lexically different from the one proposed in the glosses DU20, DU23

f) Presence of identical Noun Phrase in two coordinated sentences which sounds redundant and should have been pronominalized as has been done in the manual story
DU22, DU24

Eventually, we recorded no case of identical utterances, 8 cases of Noticeable Differences due to our algorithm and a higher number (10 cases) of arbitrary or stylistically motivated insertion of linguistic material in the manual story. The remaining mismatches (45) are to be regarded minor or Slight Differences which should be corrected in the future by further developments of the main algorithm. Overall, on a total of 54 Sentences and 91 simple sentences or propositions, we had 63 mismatches only 8 of which had a semantic impact on the story, which amounts to less than 10% error rate.

5 Conclusion

In this paper we presented the conversion process produced by *GENLIS*, a system that generates Italian text from glosses of the Italian Sign Language (LIS). The signed text we chose is a fable, i.e. a semantically and pragmatically difficult text to generate. We described all the steps that are required to convert a vector-like representation of the multilayered annotation scheme used for transcribing signs into glosses. To complete our experiment, we did an evaluation by comparing the output of the generator to a manually written version of the story to suit stylistic requirements for fables and came to the conclusion that the result is acceptable but for a few particularly difficult utterances. Eventually we only had 8 semantically relevant mismatches over 63 as a whole. However we had to overcome a number of problematic issues at a morphological, syntactic and semantic level which were successfully solved thanks to peripheral rules executed at the end of the generation process. Future work includes improving the algorithm to generate a story which is more natural and pleasant to listen to. It shall also address separately either dialogues or narrative texts in order to produce a consistent and more generalized conversion process from glosses to spoken utterances.

References

D. Barberis, N. Garazzino, P. Prinetto, G. Tiotto, A. Savino, U. Shoaib, and N. Ahmad. 2011. Language resources for computer assisted translation from italian to italian sign language of deaf people. In *Proceedings of Accessibility Reaching Everywhere AEGIS Workshop and International Conference*, Brussel.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. in proceedings of the 11th conference of the european chapter of the association for computational linguistics. pages 313–320, Trento.

Joan Bresnan. 2002. *Lexical-Functional Syntax*. Blackwells.

Cristiano Chesi, Gianluca Lebani, and Margherita Pallottino. 2008. A bilingual treebank (ita-lis) suitable for machine translation: what cartography and minimalism teach us. *STUDIES IN LINGUISTICS*, 2:165–185.

Rodolfo Delmonte. 2000. Generating from a discourse model. In *Proceedings of the MT 2000 - MACHINE TRANSLATION AND MULTILINGUAL APPLICATIONS IN THE NEW MILLENNIUM*, pages 313–320, Exeter(UK). British Computer Society - BCS.

Rodolfo Delmonte. 2016. Expressivity in tts from semantics and pragmatics. In *Il farsi e disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio*, pages 407–427, Milano, Italy.

Rodolfo Delmonte and Dario Bianchi. 1998. Dialogues from texts: How to generate answers from a discourse model. In *Atti Convegno Nazionale AI*IA*, page 139–143.

Rodolfo Delmonte, Giacomo Ferrari, Anna Goy, Leonardo Lesmo, Bernardo Magnini, Emanuele Pianta, Oliviero Stock, and Carlo Strapparava. 1996. Ilex - un dizionario computazionale dell'italiano. In *Proceedings of 5th Convegno Nazionale della Associazione Italiana per l'Intelligenza Artificiale - AI*IA "Cibernetica e Machine Learning"*, pages 27–30, Napoli, Italy.

Rodolfo Delmonte and Emananuele Pianta. 2008. Answering why-questions in closed domains from a discourse model. In *Proceedings of Semantics in Text Processing (STEP)*, page 109–114.

Brigitte Dorner and E. Hagen. 1994. Towards an american sign language interface. *Artificial Intelligence Review*, 8:235–253.

P. Dreuw, J. Forster, Y. Gweth, D. Stein, H. Ney, G. Martinez, J. V. Llahi, O. Crasborn, E. Ormel, W. Du, T. Hoyoux, J. Piater, J. M. Moya, and M. Wheatley. 2011. Signspeak – understanding, recognition, and translation of sign languages. In *Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 22–23.

E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos, and F. Goudenove. 2010. Dicta-sign: Sign language recognition, generation and modelling with application in deaf communication. In *CSLT 2010 - LREC 2010*, pages 80–83.

R. Elliott, J.R.W. Glauert, J.R. Kennaway, and I. Marshall. 2000. The development of language processing support for the visicast project. In *4th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS 2000)*, pages 101–108.

A. Gal, G. Lapalme, P. Saint-Dizier, and H. Somers. 1991. *Prolog for Natural Language Processing*. John Wiley Sons Ldt, Chinchester.

A. Gatt and E. Krahmer. 2017. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

A. Herrmann and N.-K. Pendzich. 2018. Between narrator and protagonist in fables of german sign language. *Linguistic foundation of narration in spoken and sign languages*, pages 275–308.

N. Hoiting and D.I. Slobin. 2002. *Transcription As A Tool For Understanding: The Berkeley Transcription System For Sign Language Research (BTS)*, pages 55–75. John Benjamins, Amsterdam/Philadelphia.

Leonardo Lesmo, Alessandro Mazzei, and Daniele Radicioni. 2011. Linguistic processing in the atlas project. In *Proceeding of International Workshop on Sign Language Translation and Avatar Technology(SLALT)*.

Vincenzo Lombardo, Cristina Battaglino, Rossana Damiano, and Fabrizio Nunnari. 2011. An avatar-based interface for the italian sign language. In *International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2011, June 30 - July 2, 2011, Korean Bible University, Seoul, Korea*, pages 589–594. IEEE Computer Society.

Verónica López-Ludeña, Roberto Barra-Chicote, Syaheerah Lutfi, Juan Manuel Montero, and Rubén San-Segundo. 2013. Lsespeak: A spoken language generator for deaf people. *Expert Systems with Applications*, 40:1283–1295.

C. Mellish, D. Scott, L. Cahill, D. S. Paiva, R. Evans, and M. Reape. 2006. A reference architecture for natural language generation systems. *Natural Language Engineering*, 12(01):1–34.

S. Morrissey and A. Way. 2013. Manual labour: tackling machine translation for sign languages. *Machine Translation*, 27(1):25–64.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

E. Reiter. 2010. *Natural Language Generation*, pages 574–598. Wiley-Blackwell.

D. I. Slobin, N. Hoiting, M. Anthony, Y. Biederman, M. Kuntze, R. Lindert, J. Pyers, H. Thumann, and A. Weinberg. 2001. Sign language transcription at the level of meaning components: The berkeley transcription system (bts). *Sign Language Linguistics*, 4:63–96.

D. Stein, C. Schmidt, and H. Ney. 2012. Analysis, preparation, and optimization of statistical sign language machine translation. *Machine Translation*, 26(4):325–357.

E. Sáfár and I. Marshall. 2001. The architecture of an english-text-to-sign-languages translation system. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP-02)*, pages 223–228.

Serena Trolvi and Rodolfo Delmonte. 2020. Annotating a fable in italian sign language (lis). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6025–6034, Marseille, France. European Language Resources Association.

C. H. Wu, H. Y. Su, Y. H. Chiu, and C. H. Lin. 2001. Transfer-based statistical translation of taiwanese sign language using pcfg. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6.

L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badler, and M. Palmer. 2000. A machine translation system from english to american sign language. In *Envisioning Machine Translation in the Information Future: Proceedings of the Fourth Conference of the Association for Machine Translation (AMTA-00)*, pages 293–300.

Unleashing annotations with TextAnnotator: Multimedia, multi-perspective document views for ubiquitous annotation

Giuseppe Abrami[*1], **Alexander Henlein**[*2], **Andy Lcking**[*3], **Attila Kett**[*4], **Pascal Adeberg**[*5], **and Alexander Mehler**[*6]

[*]Text Technology Lab, Goethe-University Frankfurt
[1]abrami@em.uni-frankfurt.de
[2]henlein@em.uni-frankfurt.de
[3]luecking@em.uni-frankfurt.de
[4]s7884917@stud.uni-frankfurt.de
[5]cay.ad@hotmail.de
[6]mehler@em.uni-frankfurt.de

Abstract

We argue that mainly due to technical innovation in the landscape of annotation tools, a conceptual change in annotation models and processes is also on the horizon. It is diagnosed that these changes are bound up with multi-media and multi-perspective facilities of annotation tools, in particular when considering virtual reality (VR) and augmented reality (AR) applications, their potential ubiquitous use, and the exploitation of externally trained natural language pre-processing methods. Such developments potentially lead to a dynamic and exploratory heuristic construction of the annotation process. With TEXTANNOTATOR an annotation suite is introduced which focuses on multi-mediality and multi-perspectivity with an interoperable set of task-specific annotation modules (e.g., for word classification, rhetorical structures, dependency trees, semantic roles, and more) and their linkage to VR and mobile implementations. The basic architecture and usage of TEXTANNOTATOR is described and related to the above mentioned shifts in the field.

1 Motivation

Annotation in and for computational linguistics (Gries and Berez, 2017) underwent technical and conceptual developments from XML-based annotation formats to integrated GATE (Cunningham et al., 2013) or UIMA (Götz and Suhre, 2004) frameworks (Wilcock, 2017). One reason for that development is that annotation (regardless of the annotated media such as texts, images, music, video, and so on) is bound to annotation tools, usually one annotation tool per annotation task or purpose (Cassidy and Schmidt, 2017; Dipper et al., 2004). Annotation tools are themselves subject to (technical) development (see, for instance, the annotation of rhetorical relations (Helfrich et al., 2018)). Furthermore, annotation is often part of a machine learning (ML) pipeline where machine learned applications are trained on annotated data (Rumshisky and Stubbs, 2017), so that they can later perform annotations automatically on larger data sets. This is most explicitly expressed in the MATTER/MAMA annotation model (Pustejovsky and Stubbs, 2012). In order to secure interoperability and data exchange in this dynamic landscape, annotations of linguistic phenomena (should) follow a standard (e.g. ISO, 2016).

As has been observed by Finlayson and Erjavec (2017), there are still features that are missing or only seldomly addressed in annotation tools. So further developments are here to be expected.

However, we argue that another technical and conceptual change takes place, a change that is characterised by the following, partly mutually influencing, features.

Multi-Mediality and -Perspectivity. An annotation tool trivially is a medium (for annotation). Now, as is known, for instance, from readability research, the "physical" properties of the medium *text* influence text processing: no readability difference between serif and sans serif font types has been observed (Ali et al., 2013), but they seem to differ with respect to information recall (Gasser et al., 2005). Likewise, the choice of document preparation system has an effect on the efficiency and satisfaction of the document preparer (Knauff and Nejasmic, 2015). Transferred to annotation tools, such findings evince that users may produce different results with different annotation media. Taking advantage of this effect, annotation tools

Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pages 65–75

should offer multiple views on the same data: an attribute called *Multi-Perspectivity* and realized by the tool's *Multi-Mediality*. Multi-Mediality and -Perspectivity can be realized in various ways, ranging from low-level customizable display properties to high-level exploratory means of inspecting a certain kind of data with tools/views that are developed for different data types. We conjecture both heuristic and error-reducing gains by multi-media, multi-perspective methods.

Note that multi-perspectivity is different from multimodal annotation as carried out by using video (hence the attribute 'multimodal') annotation tools such as ANVIL (Kipp, 2014) or ELAN (Wittenburg et al., 2006): while multimodality tools allow the analysis of visually recorded communication settings, multi-perspectivity tools render the same input data in various formats.

VR and AR annotations. Multi-Mediality comprises virtual reality (VR) and augmented reality (AR) as special cases. So the claims made in the previous paragraph apply here, too. However, annotating in VR or AR has some obvious repercussions on human-computer interaction (HCI, where "computer" stands for the annotation tool used). Most notably, classic HCI interfaces such as a computer mouse are replaced by locomotion or (virtual) manipulation. Again, a heuristic effect is to be conjectured, but such "immersive annotation settings" have still to be explored. A consequence is already visible, however, namely that the range of annotation objects is extended: real-world objects (AR) and the annotators' actions (VR) become potential subjects of annotations. The former is, for instance, needed in geospatial information systems (cf. Sec. 2); the latter can be used to label professional actions as learned, for instance, in virtual nurse education (Plotzky et al., 2021). We also note that VR systems are still quite new in the computational linguistics community. However, as such systems spread to all areas of human communication, people will become accustomed to their use, and the current gap between the use of traditional systems and VR will naturally disappear.

Ubiquity. Porting annotation software to mobile phones cuts any locational constraints on annotators (given a sufficient internet infrastructure). Mobile annotation probably unfold their potential when embedded into *games with a purpose* (von Ahn, 2006): annotators produce annotations "for fun" and *en passant*, when, say, being on a travel. Mobile annotations combines with AR annotations, leading to a *qualitative* (not just quantitative) change in the units of annotation.

ML for annotation (or: human-in-the-loop). The predominant annotation model conceives annotation as a means for providing data for machine learning. And annotations will surely continue to be produced and used in this way. However, the current computational annotation landscape also treads the opposing path: pre-trained ML tools are used for automatic (large-scale) annotation of documents which are then corrected by human annotators (de Castilho et al., 2019; Hemati et al., 2016). Accordingly, the role of human annotators changes from "mere" data-generators (Consten and Loll, 2012) or "two-legged meters" (Cohen, 1960) to "humans-in-the-loop" (Wagner, 2016) (i.e., a post-editing phase is interspersed at some point into the ML process, a.k.a *active learning*, Cohn et al., 1994; Settles, 2012).

Dynamics of annotation processes. It is well-known that due to an interplay of theoretical knowledge and data structure of annotation units, (linguistic) annotations exhibit a "circular" trait (Consten and Loll, 2012) – this is also reflected in the iterative design of the MAMA cycle (Pustejovsky and Stubbs, 2012). Annotation manuals and especially standardizations like the Semantic Annotation Framework (SemAF) (ISO, 2016) are means for taming this process. In fact, however, in particular the Multi-Mediality and -Perspectivity fosters the circularity of annotation processes since viewing one document from different viewpoints is a heuristic activity (cf. "Multi-Modality and -Perspectivity"). There are two consequences of this situation: Firstly, the dynamic nature of annotations is emphasized. This includes to construe annotations as parts of sequences of annotations instead of as singular tasks (cf. the argument from circularity, triggered by the mutual theoretical preconception and the actual structure of annotation data) – regardless of whether the encompassing sequence tasks are actually carried out. In other words: in designing an annotation task both (implicitly) presupposed and (potential) follow-up annotations have to be kept in mind. This is already partly reflected, for instance, in the plug-ins approach to dialogue act annotation (Bunt, 2019).

Secondly, even dynamic annotation processes

cannot afford to ignore achieved standards. On the one hand, multi-media and multi-perspective annotation tools support established schemes and ontologies. On the other hand, best practices and process standards will emerge from dynamic annotation processes.

Interim conclusion We anticipate a potential shift in thinking of and carrying out annotations, as indicated in the "Dynamics of annotation processes. This shift is driven by technological achievements mainly in the domain of VR/AR, extended pre-processing, and ubiquitous computing." Preliminary (i.e., as long as a corresponding full-blown annotation model has been developed) we refer to a system that exhibits the envisaged facilities as MUVAMP (**M**ulti-Mediality and -Perspectivity, **U**biquity, **V**R/**A**R, **M**L, **P**rocess-orientedness). Given that this is a preliminary characterization, it is obvious that no current annotation system fulfils MUVAMP. However, to making the envisaged shift happen, a precondition seems to be an annotation tool that hosts several modules (otherwise it remains unclear how multi-perspectivity is achieved). In the following we introduce TEXTANNOTATOR as a MUVAMP-oriented annotation suite for unleashing annotations along the above lines. After reviewing related approaches, we present TEXTANNOTATOR module-wise and indicate each module's role for MUVAMP.

2 Related Work

There are applications around that address some of the features outlined above. We are aware of the following ones:

- Incorporating machine learning applications into the annotation pipeline is carried out in INCEpTION (de Castilho et al., 2019) (which extends on *WebAnno* (Eckart de Castilho et al., 2016)), the commercial service *prodi.gy* (Montani and Honnibal, 2018) and the TEXTIMAGER (Hemati et al., 2016) (the latter also underlies the present work).
- Annotation in virtual reality is implemented by means of a note taking facility in (industrial) VR environments (Clergeaud and Guitton, 2017). VR visualisations have also been used in the study of multimodal referring expressions (Pfeiffer, 2012). In Wither et al. (2009), an annotation taxonomy and a prototype study on outdoor augmented reality annotation is developed.
- Mobile annotation of geospatial information is made available in MobiTOP (HoeLian Goh et al., 2012). The mobile annotation of images, e.g. for social media uses, is enabled by Anguera et al. (2008).

3 TextAnnotator

TEXTANNOTATOR is a suitable candidate as a multimedia and multimodal annotation environment for UIMA documents (Götz and Suhre, 2004). The UIMA-based annotations are driven by the TEXTANNOTATOR as a RESTfull application developed in Java. Documents which are not available in UIMA can be transferred into this format by using TEXTIMAGER (Hemati et al., 2016), which provides a rich machine learning backend for automatic annotation accouting for the 'ML' component of MUVAMP. The UIMA documents are stored through the *UIMADatabaseInterface* (Abrami and Mehler, 2018) within MongoDB[1] and can be used simultaneously and collaboratively through TEXTANNOTATOR. Collaborativity and simultaneity are enabled with bidirectional information exchange via web-socket between TEXTANNOTATOR and all client systems and is an important component for ubiquitous use. In addition, the web-socket allows other annotation tools to be connected to TEXTANNOTATOR, to ensure its multimedia nature (see Sec. 4 and 5). The connection between TEXTANNOTATOR and its client systems is illustrated in Fig. 2. Annotations stored in UIMA documents are organized in different *annotation views* (AV). Each of these views contains different annotations and is related to a specific topic or user. For each annotator, a user view is created when a document is initially opened, which duplicates the original annotations. Thus, each AV shows a different perspective, state, or context on the same document. Furthermore, the different AVs allow the computation of inter-annotator agreement, which enables to assess the consistency of annotations in a project (Krippendorff, 2018), based on user permissions.

In addition, all annotations can be used independently of TEXTANNOTATOR: they can be completely downloaded for further processing. The reuse of the annotations as a basis for ML is thus customizable, depending on the needs of the particular application. In the following, we show how TEXTANNOTATOR accounts for **Multi-Mediality and -Perspectivity**, **Dynamics of annotation pro-**

[1] https://www.mongodb.com/

cesses, and **ML for annotation**. **VR and AR**, and **Ubiquitous use** are dealt with in Sec. 4 and 5, respectively.

QuickAnnotator still allows rapid annotation of named entities and words and multi-token expression in general through a simple selection of a target class and subsequent assignment when clicking on tokens (Abrami et al., 2019). To increase annotation performance, a recommendation function was implemented that allows the selected target class, based on the token's lemma, to be applied to all other tokens of the same lemma in the same document, paragraph, or current sentence. In addition, all tokens annotated by this function are marked so that annotators can easily target and post-process them.

Another function is the combination of tokens to multitokens: By now this function has been extended with the possibility to separate tokens at any position as well as the capability to correct OCR errors (see Fig. 1). This user-friendly function, which can also be executed via drag & drop, enables the correction of incorrectly recognized token boundaries which is a frequent and popular error, especially in the context of OCR recognition of texts.

As it provides basic corrections to the texts and establishes the prerequisites for future annotation processes, QUICKANNOTATOR develops into a pre-editing tool which is employed before the main annotation work is done with more specific tools such as PROPANNOTATOR or DEPANNOTATOR. Hence, the triplet of QUICKANNOTATOR, PROPANNOTATOR, and DEPANNOTATOR gives rise to multi-

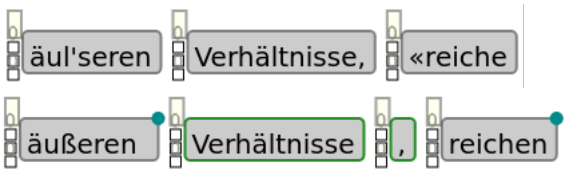


Figure 1: Tokens can be corrected as required. Firstly, incorrect token boundaries can be split using a simple key combination. In the present example this was done with the token merged with the comma (green border). Secondly, OCR errors can be corrected in QUICKANNOTATOR by clicking on the corresponding tokens. In this way the original text is not changed, but a correction token is generated, which is placed on top of the affected tokens. A corrected token is visualized with a green dot in the upper right corner. By moving the cursor over it, the original token is shown.

perspectivity (in particular if annotation files are preprocessed in such a way that they contain annotation layers according to the corresponding tools' specification).

PropAnnotator uses relations adopted from the SemAF standard (ISO, 2014a) as well as from PropBank (Palmer et al., 2005) (the latter can be mapped onto the former (Ide et al., 2017, 133)). To this end, we converted the structures defined in these standards into a UIMA type system[2]. Since the last presentation of PROPANNOTATOR in Abrami et al. (2019), significant improvements and new features have been implemented. The underlying data model allows for annotating a wide range of relations a subset of which is available in the current web interface of PROPANNOTATOR:

- Argument and modifier relations (following PropBank);
- Time relations (temporal entities and TLinks from ISOTimeML (Pustejovsky, 2017; ISO, 2012));
- Spatial-Relations (Qualitative Spatial Links (QSLinks) from ISOSpace (Pustejovsky et al., 2011; ISO, 2020a));
- A few custom extensions (for example, labeling idiomatic expressions and separated verb particles in German).

These relations are used in order to carry out semantic role labeling. Regarding semantic role labeling, the annotation of functional roles (argument and modifier relations) depends on the *sense* of the verb heading the corresponding sentence (Levin, 1993). In support of this view, PROPANNOTATOR complements semantic role annotation with verb sense annotation. Verb senses are distinguished according to PropBank's frameset lexicon[3], GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) as well as – for evaluation purposes – E-Valbu (Schumacher et al., 2004). The cross-language mixture of sense inventories is due the fact that the main language of actual annotation documents is German, but the majority of (large-scale) resource has been developed in and for English. Hence, PROPANNOTATOR provides an annotation-based mapping between English and German verb sense. Since this mapping involves translation issues, the small but hand-crafted verb

[2] https://github.com/texttechnologylab/UIMATypeSystem

[3] http://verbs.colorado.edu/propbank/framesets-english-aliases/

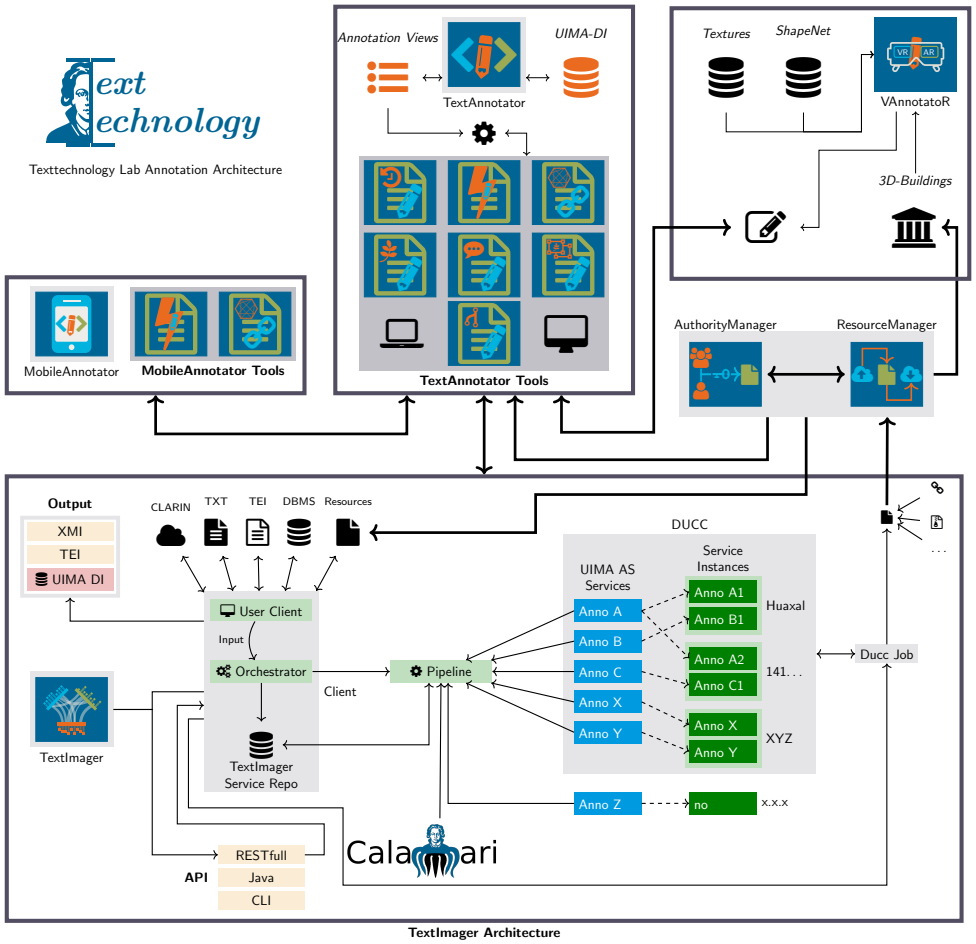


Figure 2: Illustration of the interrelationships and communication routes between the individual infrastructures. In the lower area the domain of TEXTIMAGER (Hemati et al., 2016) is shown, a multi-server system for automatic preprocessing of textual data based on UIMA. Several pipelines, each for different analyses, can be used to process the texts for use by TEXTANNOTATOR. Texts must exist in UIMA format in order to use these within the infrastructures. The upper architectures show the individual tools which are also considered in this paper in more detail. This simplified presentation shows the relationship between the tools, which all use TEXTANNOTATOR as a core service. The different annotation environments, MOBILEANNOTATOR, TEXTANNOTATOR and VANNOTATOR, are located in the upper area. Being the center of all manual annotation processes, TEXTANNOTATOR enables the use of TEXTIMAGER and thus to automatize parts of the annotation process. Each tool is directly or indirectly connected to the *ResourceManager* and *AuthorityManager* (Gleim et al., 2012) in order to manage the annotation of documents. All documents managed in ResourceManager are database objects manageable by the *UIMA database interface*. This usage takes place entirely within TEXTANNOTATOR. All tools that want to perform or use UIMA-based annotations are connected to TEXTANNOTATOR in order to subsequently use all implemented functions. Calamari, shown in the bottom region, is a Blazegraph (`https://blazegraph.com/`) implementation (still under development) for maintaining various ontologies within the TEXTANNOTATOR/TEXTIMAGER infrastructure.

sense inventory of E-Valbu is built-in as a ground truth standard of comparison.

DepAnnotator is the newest tool designed for visualization and annotation of dependency structures in texts. Based on different dependency tag sets (derived from TIGER (Brants et al., 2004), respectively NEGRA (Skut et al., 1997), and Universal Dependencies (de Marneffe et al., 2014)) existing dependencies can be deleted and new ones can be created. In addition, as with all tools, it is also possible to manually annotate texts without pre-annotated dependency information, which is illustrated in Fig. 4.

4 MobileAnnotator

To remove the binding of the annotation situation to desktop sessions, so to speak, to enable annotations in mobile contexts, quasi **ubiquitously** (whether in sitting, standing or walking

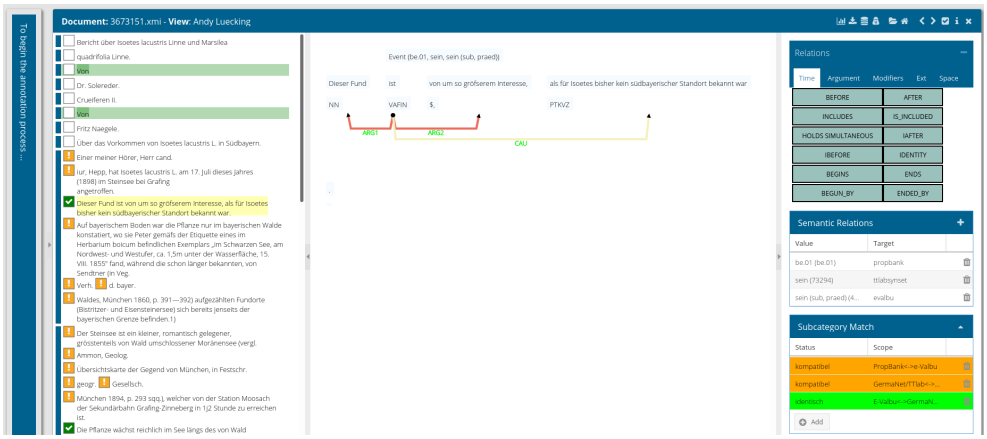


Figure 3: Visualization of PROPANNOTATOR: On the left side, the sentences of the document are displayed, which can be transferred to the middle annotation area by a click. Each sentence can be assigned with a status, which documents the annotation progress: green indicates completed sentences and yellow indicates problems, which is helpful both for later evaluation and for interrupting the process. The panel on the right side shows the annotation options. The upper part of these options shows a set of relations that can also be selected by a click. Below this is a list of annotated semantic senses, which is only enabled when an event is selected. In the center of PROPANNOTATOR's interface the annotation environment is shown by visualizing the selected sentence (left) token-wise. Similar to QUICKANNOTATOR, multitokens can be created and tokens can be separated via drag & drop. Under each token, its part of speech is displayed; clicking on a verb turns it into an "event", which can then be sense-disambiguated. Tokens can be linked semantically by drag & drop, based on the selection of a corresponding relation in the options panel (right). Colors are used to distinguish between different relation types.

position of the annotator), we have developed MOBILEANNOTATOR (Adeberg, 2020). Based on Angular[4], we adapted two tools of TEXTANNOTATOR (QUICKANNOTATOR and KNOWLEDGEBASELINKER) to enable mobile access. MOBILEANNOTATOR was developed as a TEXTANNOTATOR client (see Fig. 2) using its functionality. This allows the implementation of additional functions which are not available in the browser-based version. At the same time, documents are still accessible only after user authentication and all annotations are stored in MOBILEANNOTATOR in appropriate annotation views. The the control and use of UIMA documents is thus analogous to TEXTANNOTATOR. To motivate it with concrete examples: with MOBILEANNOTATOR, train rides, waiting time at the doctor's office, at the bus stop, or anywhere else can be used for annotation tasks. Mobile annotations, of course, attain ubiquity.

5 VAnnotatoR

VR-based annotation is provided by VANNOTATOR, a UIMA-based annotation environment implemented in Unity3D[5]. Since VANNOTATOR (Spiekermann et al., 2018) is also based on TEXTANNOTATOR (see Fig 2), its annotations can be further processed with any other annotation media of TEXTANNOTATOR.

VANNOTATOR addresses a range of scenarios: visualization and interaction with historical information (Abrami et al., 2020b), annotation of texts, their interlinking with images and 3D objects, and the creation of 3D spaces enriched with hypertext functionalities (Mehler et al., 2018; Abrami et al., 2020a).

VANNOTATOR is currently extended to include SemAF-related functionalities. A pilot study of this extension is presented in (Henlein et al., 2020). The main focus is on the annotation of spatial relations (IsoSpace, Pustejovsky et al., 2011; ISO, 2020b), semantic roles (SrLinks, ISO, 2014a) and coreference relations (MetaLinks, ISO, 2014b). This is done to generate text-to-scene data, which in turn is used to train ML systems. Fig. 6 exemplifies this sort of annotation data. In this example, we take advantage of the spatial capabilities of VR to automate as many spatial annotations as possible. That is, whenever the annotator arranges objects in virtual space based on their description in the underlying text, a subset of the relationships of the objects implied by this arrangement is explicitly annotated by the system itself. This concerns objects that are implicitly or explicitly involved in

[4] https://angular.io/
[5] https://unity.com/

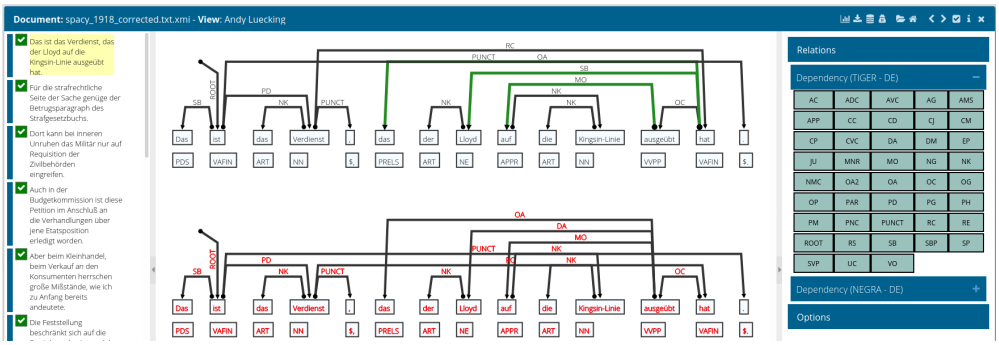


Figure 4: The interface of DEPANNOTATOR is similar to that of PROPANNOTATOR. On the left one finds the sentences to be annotated (together with their annotation status). The panel on the right displays selectable options subdivided into tagsets. Two visualizations of the focal sentence are displayed in the center of the window: the lower sentence shows the dependency tree created by the parser selected in TEXTIMAGER; the upper sentence shows its correction. Green lines encode selected dependency relations created by the human annotator. With DEPANNOTATOR it is possible to visually compare automatically created dependency trees and their corrections. DEPANNOTATOR additionally contains statistics for automatic comparison of such trees.

this description. In addition to placing entities in virtual space, the annotators' movements and gestures could be used in the future for this purpose. In a nutshell, VANNOTATOR meets the MUVAMP requirement for using VR and AR for the purpose of multimedia and multi-perspective annotation.

6 Application usage

The tools described so far are being used in various lectures and qualification work (e.g. Kühn (2018); Smaji (2020); Kett (2020); Lööck (2020)) to automatically validate annotated documents (human in the loop) or to gain new perspectives on annotated documents. In particular, TEXTANNOTATOR is used as a browser-based suite for the correction of automatic annotations generated with TEXTIMAGER.

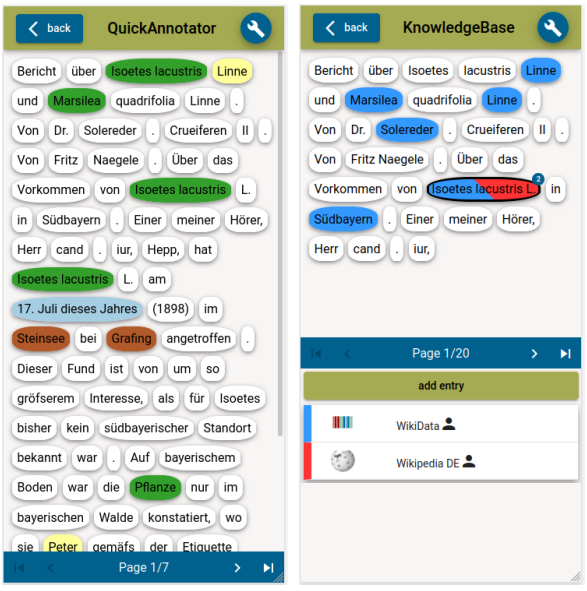


Figure 5: An excerpt from two annotation tools of MOBILEANNOTATOR. Left: adaptation of QUICKANNOTATOR; right: adaptation of KNOWLEDGEBASELINKER. By selecting a token, it can be annotated. A longer activation of tokens enables the creation of multitokens.

In addition, TEXTANNOTATOR is used in the biodiversity project *BioFID* (Driller et al., 2020). This project is concerned with the semantic indexing of historical biodiversity texts. For this purpose, TEXTANNOTATOR is used to annotate texts in order to perform various linguistic analyses. Within the BioFID project, 79,813 "net" annotations[6] have been produced using QUICKANNOTATOR (Lücking et al., 2021). "Net" means the following: since within BIOfid documents are annotated by more than one annotator for the sake of assessing inter-rater agreement, one and the same annotation unit may receive a label repeatedly but from different annotators. The net count ignores such reduplications and only takes unique labels into account.

These numbers show that many annotations can be performed by different users in a very short time (for a user evaluation of one of TEXTANNOTATOR's modules, TREEANNOTATOR, see Helfrich et al., 2018, Sec. 3). At the same time, all annotations are available in a uniform and portable format, which ultimately simplifies external processing and reuse, e.g., for ML tasks.

The combination of a large number of different annotation functions (at the word, sentence, or text

[6] as of 27th October 2020

Figure 6: Annotation example for the sentence: *He took the keys from the table and went to his car*. Left: VR view; Right: Rendering-View. In the VR annotation view one can see in yellow the *Qualitative Spatial Links* (QSLinks) and in red the *Orientation Links* (OLinks). The QSLinks and OLinks are mostly generated automatically. The thick gray line in both views represents the EventPath (here: key in hand, person to car).

level) that provide *multiple annotation perspectives* on the same text, as well as the *multimedia bandwidth* that comes with them, is, to our knowledge, currently unique in the field of annotation of natural language texts.

7 Future Work

Currently, not all annotation features available through TEXTANNOTATOR can be used by downstream tools (e.g. MOBILEANNOTATOR). To enable full ubiquitous use, different approaches for the different media (VR, AR, mobile devices) are required. In particular, we will consider the possibilities and limitations of AR systems and the extent to which they can be used for annotation purposes. While the available hardware is still very limited (price, availability, technical features, ...), in the near future it will become available to the general public, similar to VR. Furthermore, in addition to the extension of VANNOTATOR's *RoomBuilder* according to the SemAF standards, an annotation environment for TEXTANNOTATOR and MOBILEANNOTATOR is planned. This extension should make it possible to pre-annotate texts at home, in the office, or on the road, to largely complete their annotation in VR – also with regard to implied annotations – and to correct and refine the results later with conventional 2D interfaces if necessary. Insofar as these annotations refer to artifacts that are visible or even traversable in reality (e.g. streets, houses, squares), this multimedia annotation process can be significantly enriched by AR functionalities, since the direct view of the objects to be annotated can compensate for inadequacies of their representation in VR.

Thanks to the large number of tools in TEXTANNOTATOR, a wide range of annotation tasks can be addressed. However, since it is inefficient in the long run to develop tools with reference to specific annotation requirements, a more dynamic approach that simplifies the planning of annotation projects suggests itself. To meet this requirement, TEXTANNOTATOR is being further developed as a tool for modeling annotation models and corresponding annotation tools. Furthermore, it is planned to publish TEXTANNOTATOR via GitHub.

8 Conclusion

We introduced the concept of MUVAMP (**M**ultimediality and -perspectivity, **U**biquity, **VR/AR**, **ML**, **P**rocess-orientation) and argued how TEXTANNOTATOR and the annotation tools around it meet this concept. Reflecting on and studying MUVAMP, and devising corresponding annotation models is still a desideratum for computational linguistics. The increasing complexity of annotation tasks and their representation in tools in order to be able to use them collaboratively and simultaneously in a homogeneous annotation environment at best. In addition, enabling annotators to use multi-perspective multimedia annotation tools is an area where established best practices do not yet exist. In order to contribute to this research perspective, we have presented the latest developments of TEXTANNOTATOR and outlined future development steps.

In conclusion, it is our strong interest to discuss and also establish with the research community a new and more innovative way in the implementation of annotation processes. For this purpose, not only concepts and procedures are necessary, but also adequate and flexible software solutions – such as TEXTANNOTATOR.

References

Giuseppe Abrami, Alexander Henlein, Attila Kett, and Alexander Mehler. 2020a. Text2SceneVR: Generating hypertexts with vannotator as a pre-processing step for text2scene systems. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 177186, New York, NY, USA. Association for Computing Machinery.

Giuseppe Abrami and Alexander Mehler. 2018. A uima database interface for managing nlp-related text annotations. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference, May 7 - 12*, LREC 2018, Miyazaki, Japan.

Giuseppe Abrami, Alexander Mehler, Andy Lcking, Elias Rieb, and Philipp Helfrich. 2019. TextAnnotator: A flexible framework for semantic annotations. In *Proceedings of the Fifteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation, (ISA-15)*, ISA-15.

Giuseppe Abrami, Alexander Mehler, Christian Spiekermann, Attila Kett, Simon Lööck, and Lukas Schwarz. 2020b. *Educational Technologies in the area of ubiquitous historical computing in virtual reality*. Taylor & Francis.

Pascal Adeberg. 2020. MobileAnnotator: an App for TextAnnotator. bachelor's thesis, Institute of Computer Science and Mathematics, Text Technology Lab, Johann Wolfgang Goethe-Universitt, Frankfurt, Germany. Original title: MobileAnnotator: eine App für den TextAnnotator.

Luis von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Ahmad Zamzuri Mohamad Ali, Rahani Wahid, Khairulanuar Samsudin, and Muhammad Zaffwan Idris. 2013. Reading on the computer screen: Does font type have effects on web text readability?. *International Education Studies*, 6(3):26–35.

Xavier Anguera, JieJun Xu, and Nuria Oliver. 2008. Multimodal photo annotation and retrieval on a mobile phone. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, page 188194, New York, NY, USA. Association for Computing Machinery.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Journal of Language and Computation*, 2:597–620.

Harry Bunt. 2019. Plug-ins for content annotation of dialogue acts. In *Proceedings of the Fifteenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, ISA-15, pages 33–45.

Steve Cassidy and Thomas Schmidt. 2017. Tools for multimodal annotation. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 209–227. Springer Netherlands, Dordrecht.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan. The COLING 2016 Organizing Committee.

Richard Eckart de Castilho, Nancy Ide, Jin-Dong Kim, Jan-Christoph Klie, and Keith Suderman. 2019. Towards cross-platform interoperability for machine-assisted annotation. *Genomics & Informatics*, (2).

D. Clergeaud and P. Guitton. 2017. Design of an annotation system for taking notes in virtual reality. In *2017 3DTV Conference: The True Vision – Capture, Transmission and Display of 3D Video*, 3DTV-CON, pages 1–4.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine learning*, 15(2):201–221.

Manfred Consten and Annegret Loll. 2012. Circularity effects in corpus studies – why annotations sometimes go round in circles. *Language Sciences*, 34(6):702–714.

H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. 2013. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput Biol*, 9(2).

Stefanie Dipper, Michael Götze, and Manfred Stede. 2004. Simple annotation tools for complex annotation tasks: an evaluation. In *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*, LREC 2004, pages 54–62.

Christine Driller, Markus Koch, Giuseppe Abrami, Wahed Hemati, Andy Lücking, Alexander Mehler, Adrian Pachzelt, and Gerwin Kasperek. 2020. Fast and easy access to Central European biodiversity data with BIOfid. In *Biodiversity Information Science and Standards*, volume 4 of *BISS*.

Mark A. Finlayson and Tomaž Erjavec. 2017. Overview of annotation creation: Processes and tools. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 167–191. Springer Netherlands, Dordrecht.

Michael Gasser, Julie Boeke, Mary Haffernan, and Rowena Tan. 2005. The influence of font type on information recall. *North American Journal of Psychology*, 7(2):181–188.

Rdiger Gleim, Alexander Mehler, and Alexandra Ernst. 2012. Soa implementation of the ehumanities desktop. In *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities 2012, Hamburg, Germany*.

T. Götz and O. Suhre. 2004. Design and implementation of the UIMA Common Analysis System. *IBM Systems Journal*, 43(3):476–489.

Stefan Th. Gries and Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 379–409. Springer Netherlands, Dordrecht.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Philipp Helfrich, Elias Rieb, Giuseppe Abrami, Andy Lücking, and Alexander Mehler. 2018. Treeannotator: Versatile visual annotation of hierarchical text relations. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference, May 7 - 12*, LREC 2018, Miyazaki, Japan.

Wahed Hemati, Tolga Uslu, and Alexander Mehler. 2016. Textimager: a distributed uima-based system for nlp. In *Proceedings of the COLING 2016 System Demonstrations*. Federated Conference on Computer Science and Information Systems.

Alexander Henlein, Giuseppe Abrami, Attila Kett, and Alexander Mehler. 2020. Transfer of isospace into a 3d environment for annotations and applications. In *Proceedings of the 16th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 32–35, Marseille. European Language Resources Association.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT - the GermaNet editing tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

D. HoeLian Goh, K. Razikin, C. Sian Lee, E. Peng Lim, K. Chatterjea, and C. Hung Chang. 2012. Evaluating the use of a mobile annotation system for geography education. *The Electronic Library*, 30(5):589–607.

Nancy Ide, Nicoletta Calzolari, Judith Eckle-Kohler, Dafydd Gibbon, Sebastian Hellmann, Kiyong Lee, Joakim Nivre, and Laurent Romary. 2017. Community standards for linguistically-annotated resources. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, chapter 4, pages 113–165. Springer Netherlands, Dordrecht.

ISO. 2012. Language resource management – Semantic annotation framework (SemAF) – Part 1: Time and events (SemAF-Time, ISO-TimeML). Standard ISO 24617-1:2012.

ISO. 2014a. Language resource management – Semantic annotation framework (SemAF) – Part 4: Semantic roles (SemAF-SR). Standard ISO/IEC TR 24617-4:2014.

ISO. 2014b. Language resource management Semantic annotation framework (SemAF) Part 7: Spatial information (ISO-Space). Standard ISO/IEC TR 24617-7:2014, International Organization for Standardization.

ISO. 2016. Language resource management – Semantic annotation framework (SemAF) – Part 6: Principles of semantic annotation (SemAF principles). Standard ISO 24617-6:2016(E).

ISO. 2020a. Language resource management – Semantic annotation framework – Part 7: Spatial information. Standard ISO 24617-7:2020.

ISO. 2020b. Language resource management Semantic annotation framework (SemAF) Part 7: Spatial information (ISO-Space). Standard ISO/IEC TR 24617-7:2020.

Attila Kett. 2020. text2city: Spatial visualization of textual structures. bachelor's thesis, Institute of Computer Science and Mathematics, Text Technology Lab, Johann Wolfgang Goethe-Universitt, Frankfurt, Germany. Original title: text2city: Räumliche Visualisierung textueller Strukturen.

Michael Kipp. 2014. ANVIL: A universal video research tool. In Jacques Durand, Ulrike Gut, and Gjert Kristofferson, editors, *Handbook of Corpus Phonology*, chapter 21, pages 420–436. Oxford University Press, Oxford, UK.

Markus Knauff and Jelica Nejasmic. 2015. An efficiency comparison of document preparation systems used in academic research and development. *PLOS ONE*, 10(4).

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4 edition. SAGE Publications.

Vincent Roy Kühn. 2018. A gesture-based interface to VR. bachelor's thesis, Institute of Computer Science and Mathematics, Text Technology Lab, Johann Wolfgang Goethe-Universitt, Frankfurt, Germany.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Simon Lööck. 2020. Distributed annotation in virtual reality. bachelor's thesis, Institute of Computer Science and Mathematics, Text Technology Lab, Johann Wolfgang Goethe-Universitt, Frankfurt, Germany.

Andy Lücking, Christine Driller, Giuseppe Abrami, Adrian Pachzelt, Manuel Stoeckel, and Alexander Mehler. 2021. Multiple annotation for biodiversity. Developing an annotation framework among biology, linguistics and text technology. *Language Resources and Evaluation*. Accepted with minor revisions.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC'14, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alexander Mehler, Giuseppe Abrami, Christian Spiekermann, and Matthias Jostock. 2018. VAnnotatoR: A framework for generating multimodal hypertexts. In *Proceedings of the 29th ACM Conference on Hypertext and Social Media*, Proceedings of the 29th ACM Conference on Hypertext and Social Media (HT '18), New York, NY, USA. ACM.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*, to appear.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71106.

Thies Pfeiffer. 2012. Using virtual reality technology in linguistic research. In *2012 IEEE Virtual Reality Workshops*, VRW '12, pages 83–84.

Christian Plotzky, Ulrike Lindwedel, Michaela Sorber, Barbara Loessl, Peter Knig, Christophe Kunze, Christiane Kugler, and Michael Meng. 2021. Virtual reality simulations in nurse education: A systematic mapping review. *Nurse Education Today*, 101.

James Pustejovsky. 2017. ISO-TimeML and the annotation of temporal information. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 941–968. Springer Netherlands, Dordrecht.

James Pustejovsky, Jessica L Moszkowicz, and Marc Verhagen. 2011. ISO-Space: The annotation of spatial information in language. In *Proc. of the Sixth Joint ISO-ACL SIGSEM Workshop on ISA*, pages 1–9.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly, Sebastopol, CA.

Anna Rumshisky and Amber Stubbs. 2017. Machine learning for higher-level linguistic tasks. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 333–351. Springer Netherlands, Dordrecht.

Helmut Schumacher, Jacqueline Kubczak, Renate Schmidt, and Vera de Ruiter. 2004. *VALBU - Valenzwörterbuch deutscher Verben*. Narr, Tübingen.

Burr Settles. 2012. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLP-97, Washington, DC.

Alen Smaji. 2020. Development and testing of an interactive 3d city model using the example of the local public transport network of the city of frankfurt. bachelor's thesis, Institute of Computer Science and Mathematics, Text Technology Lab, Johann Wolfgang Goethe-Universitt, Frankfurt, Germany. Original title: Entwicklung und Erprobung eines interaktiven 3D-Stadtmodells am Beispiel des Personennahverkehrsnetzwerks der Stadt Frankfurt.

Christian Spiekermann, Giuseppe Abrami, and Alexander Mehler. 2018. VAnnotatoR: a gesture-driven annotation framework for linguistic and multimodal annotation. In *Proceedings of the Annotation, Recognition and Evaluation of Actions (AREA 2018) Workshop*, AREA.

S. Wagner. 2016. Natural language processing is no free lunch. In Tim Menzies, Laurie Williams, and Thomas Zimmermann, editors, *Perspectives on Data Science for Software Engineering*, pages 175–179. Morgan Kaufmann, Boston.

Graham Wilcock. 2017. The evolution of text annotation frameworks. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 193–207. Springer Netherlands, Dordrecht.

Jason Wither, Stephen DiVerdi, and Tobias Höllerer. 2009. Annotation in outdoor augmented reality. *Computers & Graphics*, 33(6):679–689.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC 2006, pages 1556–1559.

Author Index