German SRL: Corpus Construction and Model Training

Maxim Konca, Andy Lücking, Alexander Mehler[†]
Text Technology Lab, Goethe University Frankfurt
Robert-Mayer-Straße 10, 60325 Frankfurt
{konca, luecking, mehler}@em.uni-frankfurt.de

Abstract

A useful semantic role-annotated resource for training semantic role models for the German language is missing. We point out some problems of previous resources and provide a new one due to a combined translation and alignment process: The gold standard CoNLL-2012 semantic role annotations are translated into German. Semantic role labels are transferred due to alignment models. The resulting dataset is used to train a German semantic role model. With F1-scores around 0.7, the major roles achieve competitive evaluation scores, but avoid limitations of previous approaches. The described procedure can be applied to other languages as well.

Keywords: German SRL, CoNLL, label alignment, translation

1. Introduction: Why a(nother) Translation-based Approach for German SRL?

Automatic Semantic Role Labeling (SRL) (Gildea and Jurafsky, 2002) is arguably one of the most challenging tasks that *Natural Language Processing* (NLP) has yet to solve. The notion "semantic role", or "thematic role", is derived from interpreting noun phrases, clauses and adverbials in relation to the main verb of a sentence (Fillmore, 1977). Holding a thematic role amounts to being a functionally characterized teammate in the event type denoted by a sentence's main verb. This relationship can be realized in morphosyntax in various ways. Hence, there is no one-to-one correspondence between syntactic and semantic roles; both are related within a grammatical interface (Cann et al., 2000). The interface for pairing syntactic arguments with thematic roles (in generative or constraint-based grammar) is known as *linking theory* (Davis et al., 2021). For the purposes of SRL within NLP, linking can be approximated by an annotation task where a sentence, or rather the syntactic tree assigned to a sentence, is mapped onto a thematic role tree. To achieve this goal, the availability of a large amount of diverse, annotated data is of utmost importance. The modest optimism regarding high resource languages, such as English, is not out of place. However, there is a lack of SRL resources for languages like German (Daza and Frank, 2020, p. 3904). The most prominent corpus that is currently used for SRL in the German language, is CoNLL-2009 (Hajič et al., 2009) (Hajič et al., 2012), where part-of-speech tags, morphological annotations and dependency structures are extended by abstract semantic role labels according to PropBank (Bonial et al., 2015); for an example sentence (ignoring part-of-speech and morphology) see Figure 1. The number of German semantic role annotations in CoNLL are summarized in Table 1. So why not just using the German CoNLL-2009 resource for German SRL? After detailed philological considerations (readers only interested in the applied methodology can jump to Section 3), we think that there is ample motivation for opting for a different approach.

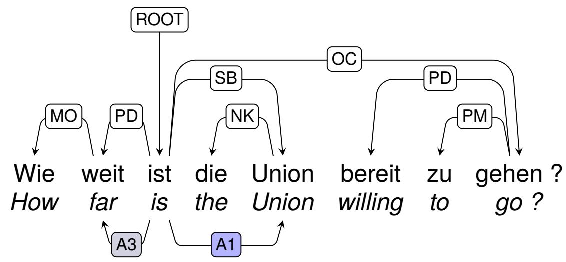

Figure 1: Example of a CoNLL-2009 annotation.

Argument	N	Argument	N
Not Annotated	284,359	A4	476
PRED	18,538	A5	222
A0	14,165	A6	63
A1	13,744	A7	77
A2	4,240	A8	49
A3	2,110	A9	7

Table 1: Number of annotations in the CoNLL-2009 dataset.

To begin with, let us have a close look at the example displayed in Figure 1, which shows the CoNLL annotation of the question *Wie weit ist die Union bereit zu gehen?* 'How far is the Union willing to go?' (where "Union" refers to the Christian Democratic Union (CDU), a political party of Germany).

[†] Alphabetical order. All authors contributed equally.

LREC-COLING 2024, pages 7717–7727

The labeled links above the string show the dependency annotation according to the dependency conversion (Seeker and Kuhn, 2014) of the TIGER scheme (Brants et al., 2004). The labeled links below the string indicate the thematic A1 and A3 arguments. It is enlightening to have a closer look at the annotation. The finite form *ist* ('is', *to be*) is the root which selects the noun of the noun phrase *die Union* 'the Union' as subject (SB), where the definite article is part of the so-called noun kernel (NK). The *Wh*-word modifies (MO) the adjective *weit* 'far', which in turn is predicative (PD) of the root. The verb *gehen* 'to go' is annotated as the copula's clausal object (OC), which, due to the infinitive construction, is built with the root morphological particle *zu* 'to' (PM). The clausal object in turn is modified by a predicative use of *bereit* 'ready'. On the thematic level, the Union is taken to fulfill the A1 relation (i.e., the patient role)[1] with regard to the root verb, while the *Wh*-phrase is an A3 argument (that is, starting point, benefactive, or attribute).

Much is wrong with this annotation. Above all, the main predicate of the sentence is *bereit sein* 'to be prepared' or 'to be willing' instead of *weit sein* 'to be far'; That is, *bereit* 'ready' has to be the PD of the root copula instead of *weit* 'far'. The actual main predicate (i.e., *bereit sein*) furthermore fails the derivation test: it can not be traced back to an active voice construction such as **Jemand bereitet die Union zu gehen*. Hence, it should be regarded as a predicational passive, not a statal one (Dudenredaktion, 1998, §322). The subjects of predicational passives, however, are likely to be "real" agents, holding role A(RG)0. This is also the way how PropBank proceeds: the predicate *to be willing* is the adjectival relation of sense *will.02* whose subject gets the agentive *desirer* role.[2] Hence, a semantic role annotation would take *bereit* 'willing' as the role-carrying predicate, *the Union* as A0, *to go* as A1, and *how far* as EXT(ent). We want to emphasize that CoNLL-2009 comprises a lot of good annotations, and just a few wrongly annotated examples are of course not the reason for refraining from using the German semantic role annotations from CoNLL-2009. We point out such an example because it causes difficulties even for large language models (LLMs; *pace* Bornheim et al., 2023).[3] Therefore, SRL cannot (yet?) be passed to LLMs; model training on annotated resources is still a valid and useful approach. But the CoNLL-2009 resource suffers from more systematic issues in this respect.

To see why we decided against using CoNLL-2009, some "resource philology" is in order. Recall first that PropBank only uses arguments numbered from ARG0 to ARG5, and in addition to that acknowledges various modifiers (e.g., for temporal or causal clauses). So where do the arguments A0 to A9 used in the German semantic role partition in CoNLL-2009 (see also Table 1) come from? CoNLL gets the German data from the SALSA corpus (Hajič et al., 2009, p. 12). SALSA in turn uses a verb-by-verb frame annotation (Burchardt et al., 2006, Sec. 3). In contrast to PropBank, frame elements (roles within frame) are frame-specific, hence there are plenty of frame-based roles. For instance, the agent of *hit* is assigned the AGENT role, its patient is an IMPACTEE. The agent of *give*, in contrast, is assigned the DONOR role, and the direct and indirect objects receive the roles THEME respectively RECIPIENT. In sum, SALSA used 628 different frames (252 FrameNet frames, and 373 newly created proto-frames) (Burchardt et al., 2006, p. 3). How does CoNLL-2009 get from frame elements to PropBank roles? The procedure is as follows (Hajič et al., 2009, p. 12): (i) In every frame, the roles AGENT, ACTOR, and COMMUNICATOR have been replaced by A0. (ii) In every frame, the roles THEME, EFFECT, and MESSAGE have been replaced by A1. The replacements in steps (i) and (ii) are percolated through the hierarchy, but involve occasional manual replacements. (iii) "Then, we used frequency and grammatical realization information to map the remaining roles onto higher-numbered Arg roles." (Hajič et al., 2009, p. 12). As a consequence of this procedure, the arguments A0 and A1, which represent prototypical agents respectively patients, have a clear interpretation. The status of the remaining arguments of the German semantic role annotation when conceived as PropBank roles is unclear at best. This is particularly true for A6 to A9, as these arguments have no direct counterpart in PropBank. In summary and in view of the fact that German SRL has not been part of later CoNLL editions, these are all reasons to try to develop a new approach based on a later CoNLL edition.

Accordingly, in this paper we present an automatically annotated German SRL corpus that is

[1] This is in fact another feature of CoNLL-2009: only heads, not complete phrases, are assigned thematic roles – this has repercussions to the evaluation described in section 4.

[2] See `http://verbs.colorado.edu/propbank/framesets-english-aliases/will.html`, last visited on 31st May 2023.

[3] ChatGPT 3.5 (default), for instance, struggled on 18th October 2023: USER: *Create a semantic role annotation according to PropBank for the following English sentence.* CHATGPT: *Of course, I'd be happy to help with that. Please provide the English sentence you'd like to be annotated with PropBank roles.* USER: *The sentence is: "How far is the Union willing to go?".* CHATGPT [shortened]: PRED=*is*, A0=*the Union*, A1 (extent)=*how far*, A2 (manner)=*willingly*, A3 (purpose)=*to go*. Same for the original German sentence.

based on the English CoNLL-2012 data (Pradhan et al., 2012) and extended by bilingual pre-aligned corpora. We review work involving SRL for the German language in section 2. The creation of the translation-based resource is described in section 3. Results are discussed in section 4. We conclude in section 5.

2. Related work

The majority of SRL applications involving the German language rely on CoNLL-2009 mainly as a benchmark for evaluation. Björkelund et al. (2009), for instance, developed a multi-modular pipeline of classifiers, that independently performed predicate disambiguation, argument identification, and argument classification. The POLYGLOT system (Akbik and Li, 2016) used English semantic role labels as universal labels and projected them to other languages, including German. Cross-View Training (Clark et al., 2018) was used by Cai and Lapata (2019a) to train a recurrent neural network for English, Chinese, German, and Spanish, that is designed to benefit from an abundance of unlabeled data to improve its performance, thus providing a way of reducing dependency on the annotated data. Cai and Lapata (2019b) exploited dependency labels without dependency parses to train an LSTM on two auxiliary tasks: (i) predicting the dependency label of a word; (ii) and predicting whether the word is directly connected to the predicate. Later, the authors proposed a method based on multilingual word embeddings (Cai and Lapata, 2020), which only makes use of semantic role annotations in the source language, raw text in the form of a parallel corpus, and an LSTM-based semantic role labeler. A language-agnostic baseline – i.e. a SRL model that does not use morphological or syntactic information – has been developed by Conia and Navigli (2020). It can be used as a "fallback solution" for low-resource languages with sparse data (such as German). X-SRL (Daza and Frank, 2020; Daza, 2022) is the approach most similar to ours, for that reason it is used as a comparison for our results in Section 4. The authors used the annotated English CoNLL-2009 dataset and translated it into French, German, and Spanish. The original English labels are then projected onto the translated datasets using the multilingual BERT model (for cosine similarity-based alignment). The authors achieved consistent labeling across the three languages, which makes evaluation and comparison significantly easier. However, X-SRL rests on head-based SRL (e.g., in Fig. 1 only *Union* receives an annotation, not the full NP *the Union*), which induces follow-up labour and difficulties when full arguments are to be retrieved (think of, e.g., the difference between restrictive and unrestrictive relative clauses), adhering to PropBank's principle that "everything within that [syntactic] span should be encompassed by an argument label" (Bonial et al., 2015, p. 20). Hartmann et al. (2016) employed linked lexical resources to generate multilingual SRL training data: sense-level information from FrameNet, WordNet, and Wiktionary, and syntactic information from VerbNet for the argument selection are combined to "exploit role-level links between VerbNet semantic roles and FrameNet roles" (p. 199). In conclusion, two requirements are desirable: Firstly, SRL should involve syntactic spans of arguments, not just heads. Secondly, a consistent labeling between resources or languages should be applied – preferrably in terms of PropBank, which, not least due to the largest available resources, can be considered the *de facto* standard.

3. Data and Methods

To create a new German SRL resource (see Section 1 for motivation), we combined automatic translation and alignment methods – see Figure 2 for an overview. The SRL-annotated English gold standard data from CoNLL-2012 (Xavier, 2022) are translated into German (see Section. 3.1.1). Simultaneously, the English language partitions of a collection of English–German parallel corpora have automatically been annotated for semantic roles (see Sections 3.1.2 and 3.1.3). Bilingual alignment information has then been exploited to project original CoNLL role annotations to the German translation (see Section 3.2). The result has been evaluated in terms of manual corrections (see Section 3.4).

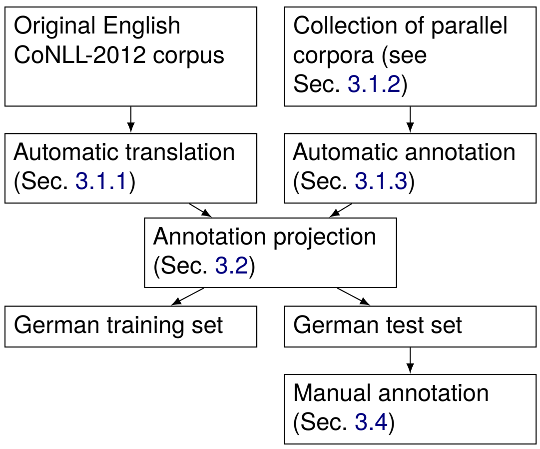

Figure 2: Workflow diagram

3.1. Data Collection

3.1.1. Translation of English CoNLL-2012 Corpus

The initial dataset utilized for this research is the English CoNLL-2012 corpus. This corpus is a

Argument	N	Argument	N
PRED	902,935	R-ARGM-LOC	2,557
ARG1	683,928	R-ARGM-TMP	1,647
ARG0	444,714	R-ARG2	1,474
ARG2	218,887	ARGM-COM	1,242
ARGM-TMP	123,516	ARGM-REC	596
ARGM-MOD	70,943	R-ARGM-MNR	363
ARGM-ADV	70,877	R-ARG3	190
ARGM-DIS	55,748	ARG5	188
ARGM-MNR	52,215	R-ARGM-CAU	138
ARGM-LOC	46,142	R-ARGM-ADV	80
ARGM-NEG	37,149	ARGA	58
R-ARG0	27,035	R-ARG4	54
R-ARG1	19,659	R-ARGM-DIR	49
ARGM-PRP	16,677	R-ARGM-PRP	46
ARG3	13,946	ARGM-PRR	26
ARGM-CAU	13,841	ARGM-PRX	26
ARGM-DIR	12,860	R-ARGM-EXT	20
ARG4	11,959	ARGM-DSP	18
ARGM-PRD	10,356	R-ARGM-COM	17
ARGM-ADJ	10,141	R-ARGM-GOL	17
ARGM-EXT	5,928	R-ARGM-MOD	3
ARGM-PNC	3,406	R-ARGM-PRD	2
ARGM-GOL	2,654	R-ARGM-PNC	2

Table 2: English CoNLL-2012 – Number of annotated arguments.

Argument	N	Argument	N
ARG1	4,255,707	R-ARGM-LOC	3,224
ARG2	1,301,532	ARGM-LVB	2,594
ARG0	1,188,718	R-ARG2	1,897
PRED	902,935	R-ARGM-TMP	1,756
ARGM-TMP	496,684	ARGM-REC	678
ARGM-ADV	472,698	R-ARGM-MNR	516
ARGM-LOC	211,836	ARGM-DSP	393
ARGM-MNR	194,706	ARG5	328
ARGM-PRP	155,939	R-ARG3	277
ARGM-CAU	144,337	ARGA	178
ARGM-PRD	85,777	R-ARGM-CAU	146
ARGM-DIS	77,342	R-ARGM-ADV	127
ARGM-MOD	71,063	R-ARGM-DIR	79
ARG3	70,864	R-ARGM-PRP	68
ARG4	53,010	R-ARG4	65
ARGM-ADJ	43,358	R-ARGM-EXT	35
R-ARG0	39,933	R-ARGM-GOL	34
ARGM-NEG	37,458	R-ARGM-COM	29
ARGM-DIR	37,099	ARGM-PRR	26
ARGM-PNC	29,620	ARGM-PRX	26
R-ARG1	23,117	R-ARGM-MOD	3
ARGM-GOL	13,380	R-ARGM-PRD	2
ARGM-EXT	11,896	R-ARGM-PNC	2

Table 3: English CoNLL-2012 – Number of annotated tokens per argument.

well-established benchmark for natural language processing tasks. The argument frequencies of the dataset used in this research are given in tables 2 and 3 – see Carreras and Màrquez (2005, p. 155) for an overview of the inventory of relation names (in addition to the PropBank roles we use "PRED" to label the semantic-role licensing predicate). The original English CoNLL-2012 corpus was translated using the state-of-the-art machine translation system DeepL.[4]

3.1.2. Collection of Parallel Corpora

We used a variety of (pre-aligned) parallel corpora from the OPUS (Tiedemann, 2012) collection (i.e., ELRA-W301 (European Language Resource Coordination 3.0), ELRC_2923 (European Language Resource Coordination 3.0, 2019b), ELRC_3382 (European Language Resource Coordination 3.0, 2019a), Salome (Poncelas et al., 2020) (Poncelas et al.), QED (Abdelali et al., 2014) (Abdelali et al.), Tilde (Rozis and Skadiņš, 2017) (Rozis and Skadins), and NewsComm (Kocmi et al., 2022) (2022 Conference on Machine Translation (WMT22) and Tiedemann) – see Table 8). The augmentation with thematically diversified data from parallel corpora leads to a more diversified dataset and makes models trained on our dataset potentially more robust.

[4] https://github.com/DeepLcom/deepl-python

3.1.3. Automatic Annotation

The collection of parallel corpora underwent an automatic annotation process. We proceeded as follows. First, we generated semantic role arguments with the model of Zhang et al. (2022)[5] (see Table 9) for English sentences, and then used the token alignments provided by the parallel corpora to transfer the arguments to the corresponding German tokens.

3.2. Annotation Projection

Both the automatically translated English CoNLL-2012 corpus and the automatically annotated parallel corpora were then merged through an annotation projection process. This step ensured that the German translations inherited the annotations from their corresponding English sentences, resulting in a dataset that retained the rich annotations of the English original while being in the German language. To align English and German tokens, we used SimAlign (Jalili Sabet et al., 2020) (see Table 9). For the argument frequencies of the projected dataset see Tables 4 and 5.

3.3. Dataset Segregation

3.3.1. German Training Set

From the merged dataset, a substantial portion was reserved to form the German training set. This

[5] A recent alternative, seq2seq model has been suggested by Přibáň and Pražák (2023).

Argument	N	Argument	N
PRED	145,634	R-ARGM-LOC	478
ARG1	128,140	R-ARGM-TMP	324
ARG0	83,760	ARGM-COM	309
ARG2	47,548	R-ARG2	244
ARGM-TMP	25,035	ARGM-REC	90
ARGM-DIS	14,442	R-ARGM-MNR	52
ARGM-ADV	13,649	R-ARG3	30
ARGM-MOD	12,019	ARGM-ADJ	24
ARGM-LOC	8,929	R-ARGM-CAU	23
ARGM-MNR	8,405	ARG5	16
ARGM-NEG	7,258	R-ARGM-ADV	11
VG	6,095	R-ARG4	11
REC	4,776	R-ARGM-DIR	10
R-ARG0	4,569	ARGA	9
R-ARG1	3,601	ARGM-LVB	5
ARG4	2,840	R-ARGM-EXT	5
ARGM-PRP	2,755	ARGM-PRR	4
ARGM-CAU	2,656	ARGM-PRX	4
ARG3	2,644	ARGM-DSP	4
ARGM-PRD	1,663	R-ARGM-PRP	4
ARGM-DIR	1,623	R-ARGM-GOL	3
ARGM-EXT	773	R-ARGM-COM	2
ARGM-GOL	617	R-ARGM-MOD	1
ARGM-PNC	612	R-ARGM-PRD	1

Table 4: Translated and aligned German CoNLL-2012 – Number of annotated arguments.

Argument	N	Argument	N
ARG1	711,435	ARGM-COM	980
ARG2	247,899	R-ARGM-LOC	761
ARG0	212,417	R-ARG2	521
PRED	146,322	R-ARGM-TMP	431
ARGM-TMP	92,561	ARGM-REC	101
ARGM-ADV	80,611	ARGM-DSP	90
ARGM-LOC	39,586	R-ARGM-MNR	87
ARGM-MNR	33,440	ARGM-ADJ	70
ARGM-CAU	25,044	R-ARG3	47
ARGM-PRP	24,562	ARG5	33
ARGM-DIS	19,168	R-ARG4	32
ARGM-PRD	13,791	ARGA	28
ARGM-MOD	12,618	R-ARGM-CAU	27
ARG3	11,890	R-ARGM-ADV	21
ARG4	11,534	R-ARGM-DIR	16
R-ARG0	7,547	R-ARGM-PRP	13
ARGM-NEG	7,522	ARGM-PRR	7
VG	6,106	ARGM-PRX	7
ARGM-DIR	5,974	R-ARGM-EXT	7
R-ARG1	5,486	ARGM-LVB	5
ARGM-PNC	5,410	R-ARGM-COM	5
REC	4,776	R-ARGM-GOL	5
ARGM-GOL	2,638	R-ARGM-PRD	2
ARGM-EXT	1,874	R-ARGM-MOD	1

Table 5: Translated and aligned German CoNLL-2012 – Number of annotated tokens per argument.

set could later be used to train various machine learning and NLP models, ensuring their compatibility and performance with the German language.

3.3.2. German Test Set

A separate subset of the merged dataset was isolated as the German test set. This would be instrumental in evaluating the performance of the trained models, providing insights into their accuracy, precision, recall, and overall efficiency. To ensure adequate argument distribution, we used stratification methods presented in (Sechidis et al., 2011) and (Szymański and Kajdanowicz, 2017).

3.4. Manual Annotation

For the final phase, the German test set underwent a rigorous manual annotation process. Two expert annotators were employed to ensure the correctness and consistency of the annotations, rectifying any discrepancies or errors that might have been introduced during the automatic processes. This step not only bolstered the reliability of the test set but also provided a gold standard against which the performance of models could be benchmarked.

Manual annotation has been carried out by making use of the PropAnnotator from the TextAnnotator collection of annotation tools (Abrami et al., 2020). The annotator agreement measured as Krippendorff's α reached respectable 0.786. The annotations were then used to measure the quality of the automatic annotation and alignment procedure (see Tables 10 and 11 – here and in the following, rows with F1-Scores of 0.7 and higher are highlighted in green, those below 0.3 in red, arguments with zero support are omitted).

Argument	N	Argument	N
PRED	1,195	ARGM-PRP	43
ARG1	1,044	ARG3	33
ARG0	621	ARGM-CAU	31
ARG2	351	ARGM-GOL	18
ARGM-TMP	205	ARGM-DIR	15
ARGM-ADV	172	ARGM-PRD	13
ARGM-MOD	139	ARGM-EXT	12
ARGM-DIS	124	R-ARGM-LOC	9
ARGM-LOC	101	ARGM-COM	6
ARGM-MNR	99	ARGM-ADJ	4
ARGM-NEG	88	ARGM-PNC	3
VG	67	ARGM-LVB	2
ARGM-REC	54	R-ARG2	2
R-ARG1	50	R-ARGM-TMP	2
ARG4	48	R-ARG3	1
R-ARG0	44	R-ARG4	1

Table 6: Translated and aligned German CoNLL-2012 – Number of arguments in the test set (Annotator 1). Stratified using (Sechidis et al., 2011) and (Szymański and Kajdanowicz, 2017).

Argument	N	Argument	N
PRED	959	ARGM-CAU	31
ARG1	828	ARG3	26
ARG0	548	ARGM-PRP	24
ARG2	291	ARGM-COM	9
ARGM-TMP	159	ARGM-GOL	7
ARGM-MOD	104	ARGM-PRD	6
ARGM-DIS	91	ARGM-DIR	5
ARGM-LOC	81	ARGM-EXT	4
ARGM-ADV	74	ARGM-PNC	4
ARGM-NEG	68	R-ARG2	2
ARGM-MNR	66	R-ARGM-LOC	2
ARG4	53	ARGM-CXN	1
ARGM-REC	50	ARGM-LVB	1
VG	46	R-ARGM-TMP	1
R-ARG1	37	R-ARG4	1

Table 7: Translated and aligned German CoNLL-2012 – Number of arguments in the test set (Annotator 2). Stratified using (Sechidis et al., 2011) and (Szymański and Kajdanowicz, 2017).

Corpus	Predicates (tokens)	Sentences
ELRA	23	12
ELRC_2923	522	284
ELRC_3382	7,248	4,262
Salome	1,337	901
QED	44,224	25,213
Tilde	8,821	6,356
NewsComm	80,854	45,460
total	143,029	82,488

Table 8: Statistics for parallel corpora.

4. Results

In order to assess the quality and efficacy of the data produced, we have trained a semantic role labeling model utilizing the state-of-the-art crfsrl algorithm (Zhang et al., 2022). The performance metrics for the test set are presented in Table 13 (for the development set see Table 12), detailing precision, recall, F1-score, and support for each argument category. Core argument roles such as ARG0 achieve a precision of 0.84, a recall of 0.68, and an F1-score of 0.75, with a support of 303 instances. ARG1 exhibits a precision of 0.71, a recall of 0.68, and an F1-score of 0.70, supported by 600 instances. ARG2 reaches a precision of 0.54, a recall of 0.48, and an F1-score of 0.51, with 201 instances in the test set. The performance varies across the modifier roles – e.g. ARGM-ADV and ARGM-CAU show moderate F1-scores of 0.21 and 0.42, respectively, with the latter having a notable

	SRL	SimAlign
F_1-score	0.86	0.81

Table 9: Model performance overview.

Argument	Precision	Recall	F1-Score	Support
ARG0	0.88	0.65	0.75	588
ARG1	0.79	0.59	0.67	976
ARG2	0.72	0.56	0.63	326
ARG3	0.75	0.60	0.67	35
ARG4	0.83	0.60	0.70	50
ARGM-ADV	0.67	0.48	0.56	159
ARGM-CAU	0.83	0.65	0.73	31
ARGM-COM	1.00	0.71	0.83	7
ARGM-DIR	0.45	0.60	0.51	15
ARGM-DIS	0.76	0.65	0.70	110
ARGM-EXT	0.50	0.30	0.37	10
ARGM-GOL	0.71	0.83	0.77	18
ARGM-LOC	0.73	0.63	0.68	90
ARGM-MNR	0.68	0.62	0.65	95
ARGM-MOD	0.88	0.58	0.70	128
ARGM-NEG	0.77	0.55	0.64	75
ARGM-PRD	0.33	0.46	0.39	13
ARGM-PRP	0.52	0.51	0.52	47
ARGM-REC	0.96	0.46	0.62	50
ARGM-TMP	0.76	0.72	0.74	198
C-ARG0	0.50	0.62	0.55	13
C-ARG1	0.45	0.62	0.52	79
C-ARG2	0.31	0.54	0.40	35
C-ARG4	0.80	0.57	0.67	7
C-ARGM-MOD	1.00	0.50	0.67	2
C-ARGM-PRP	0.00	0.00	0.00	4
R-ARG0	0.83	0.42	0.56	45
R-ARG1	0.88	0.31	0.45	49
R-ARGM-LOC	0.57	0.50	0.53	8
PRED	1.00	0.76	0.86	1,111
VG	0.98	0.81	0.88	62
micro	0.81	0.64	0.71	4,462
macro	0.52	0.44	0.45	4,462
weighted	0.83	0.64	0.71	4,462

Table 10: Performance metrics of the automatic translation and alignment evaluated on annotations of Annotator 1. Here and in the following, rows with F1-Scores of 0.7 and higher are highlighted in green, those below 0.3 in red. We omitted non-occurring arguments (zero support).

precision of 0.66. ARGM-NEG stands out with a precision of 0.69, recall of 0.84, and an F1-score of 0.76. The identification of the role-carrying predicates (PRED) reaches an impressive precision of 0.99, a recall of 0.88, and an F1-score of 0.93, for a total of 680 instances. VG abbreviates "verb group", a label which has been introduced for safety's sake to annotate the components of discontinuous verb phrases, a common phenomenon in German.[6] With an F1-score of 0.84, supported by a precision and a recall of 0.84 and 0.83, respectively, and 71 instances in the test set, it is a rather reliable label. In summary, while certain argument roles, especially core roles such as ARG0, ARG1 and specialized roles such as PRED, show commendable performance, others, especially some rarely occurring modifier roles, are difficult to apply, affecting the overall effectiveness of the system on the test set.

To assess whether the generated dataset shows improvements in annotation performance, we additionally trained an SRL model on the German

[6] For instance, the particle verb *ankommen* 'to arrive' is split in V2 sentences: *Er kommt am Bahnhof an* (He arrives at the train station). Here, *kommt* and *an* would be connected by a VG edge.

Argument	Precision	Recall	F1-Score	Support
ARG0	0.92	0.65	0.76	613
ARG1	0.80	0.62	0.70	952
ARG2	0.68	0.57	0.62	339
ARG3	0.59	0.73	0.66	30
ARG4	0.86	0.54	0.67	57
ARGM-ADV	0.61	0.65	0.62	93
ARGM-CAU	0.81	0.50	0.62	34
ARGM-COM	0.80	0.89	0.84	9
ARGM-DIR	0.39	1.00	0.56	7
ARGM-DIS	0.75	0.80	0.78	115
ARGM-EXT	0.86	0.86	0.86	7
ARGM-GOL	0.47	1.00	0.64	9
ARGM-LOC	0.73	0.66	0.69	91
ARGM-MNR	0.74	0.62	0.67	78
ARGM-MOD	0.85	0.60	0.70	121
ARGM-NEG	0.81	0.63	0.71	70
ARGM-PNC	0.57	1.00	0.73	8
ARGM-PRD	0.60	0.90	0.72	10
ARGM-PRP	0.36	0.41	0.38	32
ARGM-REC	0.97	0.58	0.72	52
ARGM-TMP	0.79	0.71	0.75	199
C-ARG0	0.42	0.62	0.50	13
C-ARG1	0.39	0.64	0.48	76
C-ARG2	0.39	0.61	0.48	46
C-ARG4	0.40	0.40	0.40	5
R-ARG0	0.88	0.56	0.69	39
R-ARG1	0.76	0.64	0.70	45
PRED	1.00	0.80	0.89	1,097
VG	0.98	0.78	0.87	51
micro	0.81	0.67	0.74	4,322
macro	0.52	0.53	0.50	4,322
weighted	0.83	0.67	0.74	4,322

Table 11: Performance metrics of the automatic translation and alignment evaluated on annotations of Annotator 2.

CoNLL-2009 dataset, [7] which we discarded in Section 1. Several key findings can be identified: the majority of the arguments (e.g., ARG0, ARG1, ARG2, and various ARGM-types) exhibit low precision and F1-scores (see Tables 14 and 15). In particular for the test set, numerous arguments show zero values across these metrics, indicating a complete lack of recognition or identification for those categories. The predicate (PRED) achieved the highest precision of 1.00, but with a low recall of 0.20, resulting in an F1-score of 0.34. ARG0, though not performing optimally, has shown relatively higher scores compared to many other arguments with a precision of 0.40, recall of 0.13, and an F1-score of 0.20. Note, however, that due to the fact that in the CoNLL-2009 dataset only the heads were annotated (see Figure 1), the recall is actually expected to be low. In summary, the model's performance on the CoNLL-2009 dataset for the majority of the arguments is suboptimal, with a few arguments exhibiting marginally better results.

Comparing results of models trained on CoNLL-2009 and CoNLL-2012 datasets, we see that the latter exhibits significantly better performance across most arguments when compared to the model trained on the CoNLL-2009 dataset. While

[7]We trained the model using both the original CoNLL-2009 and the CoNLL-2012 development sets. Nonetheless, we only report the significantly better results of the original CoNLL-2009 development set.

Argument	Precision	Recall	F1-Score	Support
ARG0	0.79	0.76	0.77	630
ARG1	0.74	0.67	0.70	1,050
ARG2	0.61	0.52	0.56	353
ARG3	0.62	0.22	0.33	36
ARG4	0.71	0.56	0.62	52
ARGM-ADV	0.53	0.28	0.36	181
ARGM-CAU	0.46	0.50	0.48	34
ARGM-COM	0.44	0.57	0.50	7
ARGM-DIR	0.46	0.38	0.41	16
ARGM-DIS	0.71	0.73	0.72	122
ARGM-EXT	0.50	0.17	0.25	12
ARGM-GOL	0.75	0.17	0.27	18
ARGM-LOC	0.53	0.56	0.54	102
ARGM-MNR	0.69	0.60	0.64	101
ARGM-MOD	0.89	0.80	0.84	142
ARGM-NEG	0.89	0.82	0.85	92
ARGM-PRD	0.60	0.23	0.33	13
ARGM-PRP	0.54	0.54	0.54	48
ARGM-REC	0.87	0.83	0.85	54
ARGM-TMP	0.65	0.77	0.70	220
C-ARG0	1.00	0.23	0.38	13
C-ARG1	0.33	0.19	0.24	79
C-ARG2	0.40	0.16	0.23	38
C-ARG4	1.00	0.14	0.25	7
R-ARG0	0.84	0.80	0.82	46
R-ARG1	0.85	0.55	0.67	51
R-ARGM-LOC	0.42	0.80	0.55	10
PRED	1.00	0.88	0.94	1,199
VG	0.91	0.93	0.92	68
micro	0.78	0.69	0.73	4,830
macro	0.45	0.35	0.37	4,830
weighted	0.77	0.69	0.73	4,830

Table 12: Prediction results of the development set for model trained on CoNLL-2012 dataset.

Argument	Precision	Recall	F1-Score	Support
ARG0	0.84	0.68	0.75	303
ARG1	0.71	0.68	0.70	600
ARG2	0.54	0.42	0.48	271
ARG3	0.00	0.00	0.00	30
ARG4	0.20	0.25	0.22	8
ARGM-ADV	0.30	0.17	0.21	206
ARGM-CAU	0.66	0.31	0.42	68
ARGM-COM	0.00	0.00	0.00	15
ARGM-CXN	0.00	0.00	0.00	22
ARGM-DIS	0.44	0.09	0.15	332
ARGM-EXT	0.50	0.11	0.18	27
ARGM-GOL	0.00	0.00	0.00	8
ARGM-LOC	0.18	0.18	0.18	136
ARGM-LVB	0.00	0.00	0.00	11
ARGM-MNR	0.35	0.28	0.31	79
ARGM-MOD	0.83	0.47	0.60	104
ARGM-NEG	0.69	0.84	0.76	37
ARGM-PRD	0.00	0.00	0.00	14
ARGM-PRP	0.47	0.33	0.39	21
ARGM-REC	0.00	0.00	0.00	47
ARGM-TMP	0.18	0.33	0.23	88
PRED	0.99	0.88	0.93	680
VG	0.84	0.83	0.84	71
micro	0.64	0.51	0.57	3,184
macro	0.26	0.20	0.22	3,184
weighted	0.63	0.51	0.55	3,184

Table 13: Prediction results of the test set for model trained on CoNLL-2012 dataset.

both models have certain arguments with zero values across precision, recall, and F1-scores, the CoNLL-2012 trained model exhibits fewer such instances, highlighting its superior capability to recognize a broader range of arguments. A notable standout in the CoNLL-2012 results is the PRED argument, with precision, recall, and F1-scores of 0.99, 0.88, and 0.93, respectively. This argument is not present in the CoNLL-2009 results.

Argument	Precision	Recall	F1-Score	Support
ARG0	0.45	0.21	0.29	630
ARG1	0.15	0.04	0.06	1,050
ARG2	0.14	0.03	0.05	353
ARG3	0.00	0.00	0.00	36
ARG4	0.00	0.00	0.00	52
ARGM-ADV	0.00	0.00	0.00	181
ARGM-CAU	0.00	0.00	0.00	34
ARGM-COM	0.00	0.00	0.00	7
ARGM-DIR	0.00	0.00	0.00	16
ARGM-DIS	0.00	0.00	0.00	122
ARGM-EXT	0.00	0.00	0.00	12
ARGM-GOL	0.00	0.00	0.00	18
ARGM-LOC	0.00	0.00	0.00	102
ARGM-MNR	0.00	0.00	0.00	101
ARGM-MOD	0.00	0.00	0.00	142
ARGM-NEG	0.00	0.00	0.00	92
ARGM-PRD	0.00	0.00	0.00	13
ARGM-PRP	0.00	0.00	0.00	48
ARGM-REC	0.00	0.00	0.00	54
ARGM-TMP	0.00	0.00	0.00	220
C-ARG0	0.00	0.00	0.00	13
C-ARG1	0.00	0.00	0.00	79
C-ARG2	0.00	0.00	0.00	38
C-ARG4	0.00	0.00	0.00	7
R-ARG0	0.00	0.00	0.00	46
R-ARG1	0.00	0.00	0.00	51
R-ARGM-LOC	0.00	0.00	0.00	10
PRED	1.00	0.30	0.47	1,199
VG	0.00	0.00	0.00	68
micro	0.52	0.11	0.19	4,830
macro	0.04	0.01	0.02	4,830
weighted	0.35	0.11	0.17	4,830

Table 14: Prediction results of the development set for model trained on CoNLL-2009 dataset.

Argument	Precision	Recall	F1-Score	Support
ARG0	0.40	0.13	0.20	303
ARG1	0.10	0.02	0.03	600
ARG2	0.15	0.02	0.04	271
ARG3	0.00	0.00	0.00	30
ARG4	0.00	0.00	0.00	8
ARGM-ADV	0.00	0.00	0.00	206
ARGM-CAU	0.00	0.00	0.00	68
ARGM-COM	0.00	0.00	0.00	15
ARGM-CXN	0.00	0.00	0.00	22
ARGM-DIS	0.00	0.00	0.00	332
ARGM-EXT	0.00	0.00	0.00	27
ARGM-GOL	0.00	0.00	0.00	8
ARGM-LOC	0.00	0.00	0.00	136
ARGM-LVB	0.00	0.00	0.00	11
ARGM-MNR	0.00	0.00	0.00	79
ARGM-MOD	0.00	0.00	0.00	104
ARGM-NEG	0.00	0.00	0.00	37
ARGM-PRD	0.00	0.00	0.00	14
ARGM-PRP	0.00	0.00	0.00	21
ARGM-REC	0.00	0.00	0.00	47
ARGM-TMP	0.00	0.00	0.00	88
PRED	1.00	0.20	0.34	680
VG	0.00	0.00	0.00	71
micro	0.48	0.06	0.11	3,184
macro	0.06	0.01	0.02	3,184
weighted	0.28	0.06	0.10	3,184

Table 15: Results of the test set for model trained on CoNLL-2009 dataset.

The micro average F1-score for the CoNLL-2012 trained model is 0.84, almost 7.5 times higher than the score of 0.11 achieved by the CoNLL-2009 model. Similarly, the macro and weighted averages for the CoNLL-2012 model are substantially higher. Training on the CoNLL-2012 dataset seems to significantly enhance the model's ability to recognize and predict a broader range of semantic roles and arguments. The CoNLL-2012 trained

Argument	%	Argument	%
ARG0	0.67	ARG0	0.62
ARG1	0.59	ARG1	0.63
ARG2	0.53	ARG2	0.49
ARG3	0.49	ARG3	0.80
ARG4	0.52	ARG4	0.49
ARGM-ADV	0.43	ARGM-ADV	0.58
ARGM-CAU	0.63	ARGM-CAU	0.47
ARGM-DIS	0.46	ARGM-DIS	0.52
ARGM-LOC	0.58	ARGM-LOC	0.63
ARGM-MNR	0.49	ARGM-MNR	0.60
ARGM-MOD	0.54	ARGM-MOD	0.65
ARGM-NEG	0.60	ARGM-NEG	0.54
ARGM-PRP	0.46	ARGM-PRP	0.39
ARGM-REC	0.00	ARGM-REC	0.00
ARGM-TMP	0.67	ARGM-TMP	0.61
C-ARG1	0.03	C-ARG1	0.01
C-ARG2	0.00	C-ARG2	0.00
PRED	0.61	PRED	0.63
R-ARG1	0.35	R-ARG1	0.31
VG	0.00	VG	0.00

Table 16: Percentage of correctly projected arguments from manually annotated data using the X-SRL model (on the left – Annotator 1, on the right – Annotator 2).

model demonstrates notably higher precision and recall across the majority of arguments compared to the CoNLL-2009 trained model. The improved aggregate metrics (micro, macro, weighted) for the CoNLL-2012 model suggest that it may generalize better to various semantic roles and contexts in the test set. We also tested the X-SRL projection pipeline of Daza and Frank (2020); Daza (2022). X-SRL does not lead to improvement on our data. Table 16 shows the percentage of arguments that were projected accurately. It has to be noted, however, that some portion of disagreement can be due to the discrepancies between original English CoNLL annotations and their manual corrections in the German language. In conclusion, training on the CoNLL-2012 dataset offers substantial advantages in terms of recognition capabilities, accuracy, and overall performance in semantic role labeling tasks.

5. Conclusions

The development of semantic role labeling models for multiple languages remains to be a challenge, mainly due to the lack of extensively annotated datasets in languages other than English. Our methodology, illustrated in Figure 2, addresses this issue by leveraging the richly annotated English CoNLL-2012 corpus. Through a series of steps – involving automatic translation, parallel corpus collection, automatic annotation, alignment, and annotation projection – we have generated German training and test sets. This strategy aims to enrich the data availability for German semantic role labeling without the need for laborious manual annotation from scratch.
Upon evaluating our model on the German test set, the results (see tables 12 and 13) are mixed.

Core argument roles, such as ARG0 and ARG1, achieve decent F1-scores, indicative of the effectiveness of the translation and annotation projection steps. Predicate (PRED) identification also shows commendable results. However, the model struggles in accurately identifying several modifier roles, with many roles showing negligible or zero precision, recall, and F1-scores. This divergence in performance highlights the complexity of semantic role labeling, especially when relying on projected annotations from another language.

Thus, our findings indicate areas for improvement. In particular modifier roles seem to be subject to nuances and intricacies of the German language that might not be fully captured through translation and projection alone. The low frequency and zero scores in several modifier roles indicate potential pitfalls in the methodology, suggesting the need for more refined translation or projection techniques, or the incorporation of manual intervention to refine annotations in challenging areas.

In sum, our study underscores the viability of using cross-lingual projection methodologies for populating semantic role annotations in languages with limited annotated resources based on the workflow developed here. Combined with the use of large language models, this approach could help to fill the gap that still exists in SRL, especially for languages that are not considered to be low resource. We plan to publish the translated German SRL resource via LDC, which distributes ONTONOTES (Weischedel et al., 2013), the source of the CoNLL datasets.

Acknowledgements

We want to thank three anonymous LREC-COLING reviewers for their valuable feedback. Funding by the *Deutsche Forschungsgemeinschaft*, grant ME 2746/5-2, is gratefully acknowledged.

6. Ethics Statement

The authors have no competing interests to declare that are relevant to the content of this article.

7. Bibliographical References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Giuseppe Abrami, Manuel Stoeckel, and Alexander Mehler. 2020. TextAnnotator: A UIMA based tool for the simultaneous and collaborative annotation of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 891–900, Marseille, France. European Language Resources Association.

Alan Akbik and Yunyao Li. 2016. Polyglot: Multilingual semantic role labeling with unified labels. In *Proceedings of ACL-2016 System Demonstrations*, pages 1–6.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena Hwang, Martha Palmer, and Nicholas and Reesem. 2015. English PropBank annotation guidelines. Technical report, Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder.

Tobias Bornheim, Niklas Grieger, Patrick Gustav Blaneck, and Stephan Bialonski. 2023. Speaker attribution in german parliamentary debates with qlora-adapted large language models.

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.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *LREC*, pages 969–974.

Rui Cai and Mirella Lapata. 2019a. Semi-supervised semantic role labeling with cross-view training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural*

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1018–1027, Hong Kong, China. Association for Computational Linguistics.

Rui Cai and Mirella Lapata. 2019b. Syntax-aware Semantic Role Labeling without Parsing. *Transactions of the Association for Computational Linguistics*, 7:343–356.

Rui Cai and Mirella Lapata. 2020. Alignment-free cross-lingual semantic role labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3883–3894.

Ronnie Cann, Claire Grover, and Philip Miller, editors. 2000. *Grammatical Interfaces in HPSG*. Studies in Constraint-Based Lexicalism. CSLI Publications, Stanford, CA.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning*, CoNLL-2005, pages 152–164.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anthony R. Davis, Jean-Pierre Koenig, and Stephen Wechsler. 2021. Argument structure and linking. In Stefan Müller, Anne Abeillé, Robert D. Borsley, and Jean-Pierre Koenig, editors, *Head Driven Phrase Structure Grammar: The handbook*, number 9 in Empirically Oriented Theoretical Morphology and Syntax, pages 315–367. Language Science Press, Berlin.

Angel Daza. 2022. *Cross-lingual Semantic Role Labeling through Translation and Multilingual Learning*. Ph.D. thesis, Department of Computational Linguistics.

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914, Online. Association for Computational Linguistics.

Dudenredaktion, editor. 1998. *Duden, Grammatik der deutschen Gegenwartssprache*, 6 edition, volume 4 of *Duden*. Dudenverlag, Mannheim and Leipzig and Wien and Zürich.

Charles J. Fillmore. 1977. Scenes-and-frames semantics. In Antonio Zampolli, editor, *Linguistic Structures Processing*, number 5 in Fundamental Studies in Computer Science, pages 55–81. North-Holland Publishing Company, Amsterdam.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, page 1–18, USA. Association for Computational Linguistics.

Silvana Hartmann, Judith Eckle-Kohler, and Iryna Gurevych. 2016. Generating Training Data for Semantic Role Labeling based on Label Transfer from Linked Lexical Resources. *Transactions of the Association for Computational Linguistics*, 4:197–213.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alberto Poncelas, Jan Buts, James Hadley, and Andy Way. 2020. Using multiple subwords to improve English–Esperanto automated literary translation quality. *arXiv preprint arXiv:2011.14190*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40.

Pavel Přibáň and Ondřej Pražák. 2023. Improving aspect-based sentiment with end-to-end semantic role labeling model.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL – multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases*, pages 145–158.

Wolfgang Seeker and Jonas Kuhn. 2014. An out-of-domain test suite for dependency parsing of German. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC, pages 4066–4073. European Language Resources Association (ELRA).

Piotr Szymański and Tomasz Kajdanowicz. 2017. A network perspective on stratification of multi-label data. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, pages 22–35, ECML-PKDD, Skopje, Macedonia. PMLR.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu, and Min Zhang. 2022. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. In *Proceedings of COLING*, pages 4212–4227, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

8. Language Resource References

2022 Conference on Machine Translation (WMT22) and Tiedemann, Jörg. *News-Commentary v16*. https://opus.nlpl.eu/News-Commentary.php.

Abdelali, Ahmed and Guzman, Paco and Sajjad, Hassan. *QCRI Educational Domain Corpus*. https://opus.nlpl.eu/QED.php.

European Language Resource Coordination 3.0. *ELRA-W0301: ELRC_403_Letter of rights for persons arrested*. https://opus.nlpl.eu/ELRA-W0301-v1.php.

European Language Resource Coordination 3.0. 2019a. *COVID-19 EU presscorner v1 dataset. Bilingual (EN-DE)*. ELRC-SHARE, 1.0.

European Language Resource Coordination 3.0. 2019b. *COVID-19 EUROPARL dataset v1. Bilingual (EN-DE)*. ELRC-SHARE, 1.0.

Hajič, Jan and Martí, Maria Antonia and Marquez, Lluis and Nivre, Joakim and Štěpánek, Jan and Padó, Sebastian and Straňák, Pavel. 2012. *2009 CoNLL Shared Task Part 1*. Linguistic Data Consortium (LDC), LDC2012T03, https://doi.org/10.35111/7y3b-fj75.

Poncelas, Alberto and Buts, Jan and Hadley, James and Way, Andy. *Salome v1*. https://opus.nlpl.eu/Salome-v1.php.

Rozis, Roberts and Skadins, Raivis. *Tilde MODEL Corpus – Multilingual Open Data for European Languages*. https://opus.nlpl.eu/TildeMODEL.php.

Weischedel, Ralph and Palmer, Martha and Marcus, Mitchell and Hovy, Eduard and Pradhan, Sameer and Ramshaw, Lance and Xue, Nianwen and Taylor, Ann and Kaufman, Jeff and Franchini, Michelle and El-Bachouti, Mohammed and Belvin, Robert and Houston, Ann. 2013. *OntoNotes Release 5.0*. Linguistic Data Consortium (LDC), LDC2013T19, https://doi.org/10.35111/xmhb-2b84, 5.0, ISLRN 151-738-649-048-2.

Xavier, Frank. 2022. *ontonotes-conll2012*. Mendeley Data, https://doi.org/10.17632/zmycy7t9h9.2, v2.