Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4605–4614
Marseille, 11–16 May 2020

Fine-grained Named Entity Annotations for German Biographic Interviews

Josef Ruppenhofer[1], **Ines Rehbein**[2], **Carolina Flinz**[3]
1. Leibniz-Institute for the German Language, R5, 6-13 , 68161 Mannheim
2. Mannheim University, Data and Web Science Group
3. Università degli Studi di Milano, Department of Studies in Language Mediation and Intercultural Communication
ruppenhofer@ids-mannheim.de, ines@informatik.uni-mannheim.de, carolina.flinz@unimi.it

Abstract

We present a fine-grained NER annotations scheme with 30 labels and apply it to German data. Building on the OntoNotes 5.0 NER inventory, our scheme is adapted for a corpus of transcripts of biographic interviews by adding categories for AGE and LAN(guage) and also adding label classes for various numeric and temporal expressions. Applying the scheme to the spoken data as well as a collection of teaser tweets from newspaper sites, we can confirm its generality for both domains, also achieving good inter-annotator agreement. We also show empirically how our inventory relates to the well-established 4-category NER inventory by re-annotating a subset of the GermEval 2014 NER coarse-grained dataset with our fine label inventory. Finally, we use a BERT-based system to establish some baselines for NER tagging on our two new datasets. Global results in in-domain testing are quite high on the two datasets, near what was achieved for the coarse inventory on the CoNLLL2003 data. Cross-domain testing produces much lower results due to the severe domain differences.

Keywords: Named Entity Recognition, spoken language, German, oral history corpora

1. Introduction

While Named Entity Recognition (NER) is typically envisioned in service of NLP tasks such as information extraction, question answering, automatic translation, etc (Jurafsky, 2000), we are interested in it also from a corpus linguistic and digital humanities perspective.

The Datenbank für Gesprochenes Deutsch (DGD; 'Database for Spoken German') that is hosted by the Leibniz Institute for the German Language is a repository of, and a platform for research on, spoken German (Schmidt, 2014). It contains a large, continuously growing collection of currently 34 variational and conversational corpora, totalling more than 4.000 hours of audiovisual material, which are used to address a wide variety of research questions. As a first step in adding a layer of shallow semantic analysis to these spoken corpora, we want to provide NER tags. Out of all the corpora in the database, we chose to begin with the ISW corpus, which contains transcripts of German-language biographic narrative interviews with Austrian-born emigrants to Israel. The ISW corpus is part of a series of three interrelated corpora (IS, ISW, ISZ) collected by Prof. Anne Betten and colleagues over more than two decades, mainly in the 1990s and the 2000s. All three corpora have been and continue to be intensely researched from various angles (Betten and Du-Nour (2004; Leonardi and Thüne (2011; Leonardi et al. (2016), *inter alia*). Our initial focus on the ISW corpus is motivated by two considerations. First, we generally want to make the data more easily accessible to researchers with a qualitative interest in the data, for instance, historians and conversation analysts. Secondly, whereas most studies of the ISW interviews so far have been limited to parts of the corpus, we want to support larger scale corpus linguistic investigations. Along these lines, Flinz (2019) and Brambilla and Flinz (2019) studied the expression of emotions in the interviews in the full ISW corpus.

The two largest datasets available for German NER, the CoNLL 2003 Shared Task dataset and the GermEval 2014 dataset, are based on newspaper and Wikipedia and news data, respectively. Our main dataset is quite different from these. First, it represents spontaneous speech, that is, it contains disfluencies, repairs and aborted utterances. In addition, the speakers also use syntactic patterns and lexical items that are much rarer in written language or sound a bit out of time to contemporary speakers of German in Germany. Second, contentwise, the corpus is focused on historical events in Austria and Germany in the 1930s as well as on Israeli and Jewish history from the 1920s to the early 2000s. And finally, whereas the CoNLL and GermEval datasets include only labels for Person, Organization, Location and Miscellaneous categories, we want to make certain additional types of information available which either are not covered at all by the other datasets, treated as subcases of the Miscellaneous class, or only partially covered by some special subclasses that GermEval provides for words related to NEs by morphological processes of compounding or derivation. Some of the classes that we use are specifically relevant for our biographic interviews are AGE, LAN(guage), and NRP (national, ethnic, religious, political or other identity).

Besides our main dataset we also apply our named entity inventory to a collection of news teaser tweets. We do this to make sure that the inventory is more generally applicable than within our domain of interest but also so we can compare the distribution of labels on our spoken interview dataset to text that is more similar to classical newspaper corpora. All our annotated data is publicly available.[1]

[1] `https://github.com/josefkr/spoken_ner_de`
`http://agd.ids-mannheim.de/isw-ner.shtml`

2. Related Work

2.1. NER on German

The CoNLL-2003 Shared Task on Language-Independent Named Entity Recognition provided a German dataset (Tjong Kim Sang and De Meulder, 2003) annotated for the traditionally used four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to one of the other three groups. The full datasetcontains 310.318 tokens in 18.933 sentences from 909 source documents. The source documents are articles from the Frankfurter Rundschau newspaper authored in 1992, which are part of LDC's ECI Multilingual Text Corpus (Linguistic Data Consortium, 1994). 7.7% of the tokens in the training set are part of a named entity.

The GermEval 2014 NER Shared Task (Benikova et al., 2014a) introduced a dataset of sentences sampled from German Wikipedia and News Corpora as a collection of citations. It contains 31.000 sentences with more than 590.000 tokens. The shared task's data was annotated following the NoSta-D guidelines (Benikova et al., 2014b), which themselves are an extension of the Tübingen Treebank guidelines (Telljohann et al., 2012) .The GermEval NER dataset contains more than 41,000 NE annotations. 7.8% of these are nested in other NEs. A typical example of this case are organization names containing place names such as [1. FC [Köln LOC] ORG] '1. FC Cologne'. NoSta-D also recognizes two NE subclasses for a) words derivationally related to the main classes (e.g. [österreichische LOCderiv] Behörde 'Austrian agency') and b) words that themselves are not NEs but which contain an NE as one of their parts ([EU-Verwaltung ORGpart] 'EU administration'). About 5.6% of the NEs in the GermEval dataset are parts of NEs concatenated with other words (part) and 11.8% are derivations. 9.3% of tokens are covered by an NE label.

2.2. Fine-grained NER

While much recent work (Ling and Weld, 2012; Gillick et al., 2014) uses fine-grained categories that are pre-defined by knowledge bases such as Freebase (Bollacker et al., 2008) or YAGO (Suchanek et al., 2007), we focus on related research that manually builds an entity set or hierarchy for domains of interest.

Fine-grained NER can mean several things. A scheme may subdivide some of the classic, large classes; attempt fine-grained coverage for general discourse; structure the entities of a specific or domain finely; or combine more than one of the foregoing aspects.

Sekine and Nobata (2004) developed an Extended Named Entity Hierarchy (ENEH) with 200 categories in three layers.[2] The ENEH intends to be domain-general and includes, for instance, paths from the `Name`-root such as Name→Facility→Line→{Railroad, Road, Canal, Water Route, Tunnel, Bridge} as well as Name→Product→Rule→{Rule_Other, Treaty, Law}. Besides the `Name` root, the inventory has roots for `Time` (e.g. Timex_other *Spring semester*) and `Numex` (numerical) expressions (e.g. Frequency *twice, five times*), which are not NEs but which are of interest for downstream applications. In fact, the first Named Entity tag set introduced by Grishman and Sundheim (1996) already included categories for percentages, time and monetary expressions.

By contrast, the work of Leitner et al. (2019) is more narrowly focused. They annotate German data from the legal domain with 19 fine-grained classes that can be mapped to 7 coarse classes.10 of the 19 fine categories are law-related and 9 are domain independent. Similarly, in work on entity recognition in traffic-related events, Schiersch et al. (2018) expand the classical 4-category NE inventory with domain general subtypes (e.g. ORG-COM(mercial) for businesses), domain-specific subtypes (e.g. Location-Stop for public transit stops) and new domain-specific top-level types (e.g. for Distance expressions).

2.3. NER on speech

As our work is focused on biographic narratives, we do not require a large and deep hierarchy of NEs, many of which would never occur in our data. On the other hand, we want more detail on some of the classic four categories as well as add some custom ones. Against this background, we take the categories used by the OntoNotes project (Weischedel et al., 2013) as a starting point. It recognizes 10 types of named entities and 7 types of what it calls values, such as MONEY and PERCENT. We add some categories and adjust and expand the definitions so they suit our overall inventory and account for linguistic facts of German as needed. The most comparable scheme to ours is the QUAERO scheme of Rosset et al. (2011; Grouin et al. (2011) that was used by the ETAPE evaluation campaign for French (Gravier et al., 2012) for NER annotation on a corpus of TV broadcast speech. It is hierarchical with 32 subcategories under 7 supercategories, whereas our scheme with 31 labels is flat. The schemes differ somewhat in where they add detail but they have many similar or overlapping categories.

3. Data

3.1. Israel corpus

The main dataset we work with is the "Israel-Korpus: Wiener in Jerusalem", or ISW corpus for short. It consists of transcriptions of biographic interviews of Israeli citizens who emigrated from Vienna, Austria, during the 1930s (Betten et al., 1995). The interviews in the ISW corpus were conducted in German and revolve around the personal and historical context of the subjects' emigration and their subsequent relationship to their city and country of origin as well as their native language, German. The version of the corpus that we use consists of 83 transcripts from 28 recording sessions involving 24 different speakers. On average each of our transcripts has 2.822 tokens, for a total of 234.271 tokens. The corpus transcripts use a literary style of transription in German standard orthography, including the use of capitalization and inter-punctuation. Some deviations from standard orthography are used to reflect notable phonetic phenomena and dialectal variants. Hesitations and disfluencies are transcribed as well. Although the data represent spoken language, and the speech of long-time emigrants in particular, the fluency of the speech is high. Due to the interview setting, the corpus contains long stretches

[2] `https://nlp.cs.nyu.edu/ene/`

Name	City	Country	Type	Tokens
Allgemeine Zeitung	Mainz	DE	regional	3.275
Handelsblatt	Düsseldorf	DE	national	3.112
Krone	Vienna	AT	national	5.794
Lausitzer Rundschau	Cottbus	DE	regional	3.198
Leipziger Volkszeitung	Leipzig	DE	regional	6.008
Mannheimer Morgen	Mannheim	DE	regional	4.859
NZZ	Zurich	CH	national	6.749
Der neue Tag	Weiden	DE	regional	8.145
Salzburger Nachrichten	Salzburg	AT	regional	2.535
SZ Kultur	Munich	DE	national	6.495
SZ Wirtschaft	Munich	DE	national	7.602
TAZ	Berlin	DE	national	8.270
Westdeutsche Allgemeine Zeitung	Essen	DE	regional	6.182
Weserkurier	Bremen	DE	regional	6.526

Table 1: Newspapers included in news tweets dataset

of speech by a single speaker and is much less interactive than spontaneous conversation, with few overlaps and interruptions. The ISW corpus is part of a series of three related corpora containing narrative biographic interviews collected by Anne Betten in the 1990s and 2000s.[3]We hope to automate NER tagging for the other two corpora, IS[4] and ISZ[5]s, using models trained on ISW.

3.2. News tweets

As a second dataset we use tweets from 13 regional and national German language newspapers published in Austria, Germany, and Switzerland (cf. Table 1 and Fig. 1). In particular, we collected data from the official accounts of the publishers as these are used to disseminate teaser texts complete with final links to full articles. These teaser tweets are often very similar to the introductory sections of news articles, which means they are informationally dense and tend to contain many named entities, especially Persons and Organizations. Overall, the twitter dataset contains 78.750 tokens (including URLs). 22.5% (17.713) of the tokens bear an NER label (17.2% if we discount the URLs.)

4. Annotation

We apply a single layer of annotation. That is, no nesting is allowed. In cases where one named entity is included in another, we label the larger span. Formally, we use the IOB2 labeling scheme. That is the first token of an NE receives the relevant label prefixed by “B-”, a non-initial token belonging to an NE receives a label with the prefix “I-”, and tokens outside of any NE are labeled as “O”.

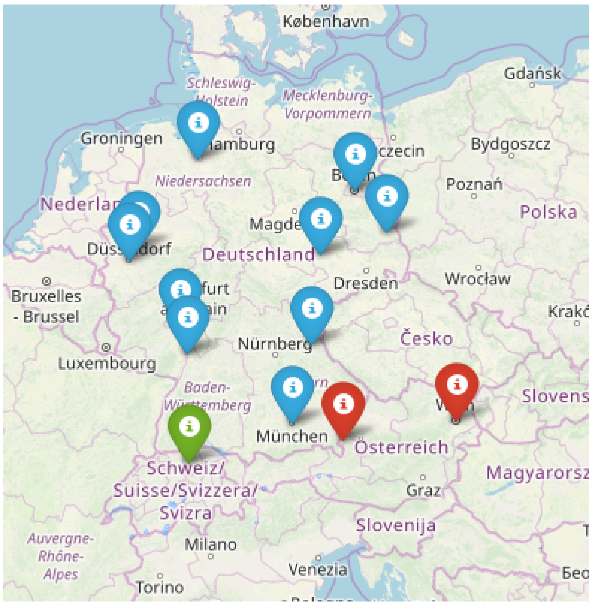

Figure 1: Geographic distribution of newspapers from which tweets were sampled

4.1. Label inventory

Our fine-grained label set is inspired by previous work that has attempted fine-grained NER annotations as well as by the discussions of subcategories of the traditional big four categories (PER, ORG, LOC and MISC), contained for instance in the annotation guidelines of the CoNLL NER Shared Task[6]. Overall, we use 30 different labels, grouped for reference into 5 supercategories, as well as a remainder category MISC for miscellaneous named entities. They are listed with brief descriptions and their number of occurrence in our two datasets in Table 2.

We note that not all of the labels that we include are named entities in a strict sense. But as noted before, since its earliest days annotations under the named entity rubric have often included other types of expressions such as time or money for application-driven purposes. Secondly, some of the labels in the inventory occur relatively rarely in the ISW data and the relevance of others may not seem particularly great for the analysis of biographic narrative interviews. Still we do use the full set in the annotation of the ISW corpus because ultimately we will want to apply NER annotations not only to the other two Israel-related corpora but to all the corpora in the Database for Spoken German. We expect that different subsets of labels will be important for different corpora.

Agentive This set of labels has to do with entities that appear as agents or protagonists in stories and reports. We include geopolitical entities here because, as administrative units, they are often important actors rather than mere spatial settings. Also note that we treat media organizations as ORG when they figure as corporate entities – e.g. *Zuckerberg is the CEO of Facebook* – but as MED when their homonymous products are referred to in contexts of users – *I posted it on Twitter*. Of special interest to us is the category NRP which we use to capture national-

[3] https://www.uni-salzburg.at/index.php?id=23142

[4] http://agd.ids-mannheim.de/IS--_extern.shtml

[5] http://agd.ids-mannheim.de/ISZ--_extern.shtml

[6] https://www.clips.uantwerpen.be/conll2003/ner/

	Abbr. label	Description	Examples	ISW corpus	News tweets
Agentive	PER•	Person names, including nick names	Schuschnigg; Greta Thunberg	925	1598
	ORG•	Organization	Mainz05; Likud	328	1760
	CREAT	Creatures, non-human	Raupe Nimmersatt	3	13
	NRP•	National, religious , political (and other identity) categories	Israeli; English	2577	667
	MED	the media products / channels of newspapers and companies like Twitter, the Jerusalem Post, etc.	[read] NY Times	13	27
	GPE•	geopolitical entities such as countries, states and cities	Austria; Vienna	3759	1944
Spatial	LOC•	locations that are not administrative units (i.e. not GPEs)	Negev, Black Sea	175	263
	ADD	Addresses in the physical world	R5 6-13, 68176 Mannheim	5	0
	URL	Addresses in virtual domains	`http://www.lrec-conf.org`	0	4181
Hum. creat.	FAC•	Facilities	Oberfalz-Kaserne [barracks]	291	305
	LAN•	Natural languages and their varieties	Hebrew; Viennese	1772	7
	LAW•	laws, ordinances, treaties etc.	Oslo Accords	15	53
	ART	works of art such as movie, song, book titles	#WhiteAlbum	93	156
	PRODUCT•	commercial products	Boeing 737 MAX	13	103
Times and events	EVT•	Events	Anschluss; Six-Day War	166	285
	PROJ	names of projects	#FridaysForFurture	10	23
	TIME•	temporal locations	now, 10 o'clock	3562	540
	DATE•	dates, a special subset of temporal locations	September 1, 1939; this month	1081	700
	DUR	durations	7 hours, four days	1380	250
	FREQ	information about frequency of events	twice; four times a day	520	50
	SORD	reference to place of repeated event in serial order	for the 2nd time	243	130
	AGE	age specificiations	10 years old	402	166
Numeric expr.	CARDINAL•	cardinal numbers	10, ten	556	397
	ORDINAL•	ordinal numbers	1.; first	384	130
	PERC•	percentages	90%; ten percent	33	20
	FRAC	fractions	3/5	22	6
	QUANT•	combinations of numbers and units of measurements	7 kilos	32	32
	RATE	distribution of one set of units over another unit	60 km/h	14	8
	MON•	amounts of money	S70.000; 3 Piaster	49	93
	SCORE	specifications of sports and other scores	6:2,6:1 [tennis]	0	21
	MISC	other NEs not covered by the above categories	Nobel Peace Prize	142	149

Table 2: Label inventory and number of instances in datasets ('•': has an OntoNotes (near)equivalent)

ity, religion, political orientation and other dimensions of identity.[7] Note that in the coarse-grained Nosta-D scheme, many of these terms are enumerated as semantic subclasses of the category MISC or to be labeled as LOCderiv, if they relate to cities, regions or countries (*Berliner.n* 'person from Berlin', *Irin.n* 'Irishwoman'). The NRP label is used both for nouns, as just seen, as well as adjectives (e.g. *westfälisch* 'related to Westphalia'; *kommunistisch* 'communist').

Spatial This group covers locations and addresses in the real and the virtual world.

Human creations The group includes categories of human created concrete and abstract things. The category for language names included here is particularly important in the context of our biographical narratives. The interviews prominently take up the issue how the interviewees experienced the change in their cultural and linguistic environment following their emigration to Israel. Before emigration, the German language and/or its Austrian variant were a key part of their personal identity and everyday life, whereas in their new life in Israel it typically receded into strictly private use with other emigrants, not least because of negative attitudes of the surrounding community towards its continued use. Note that, within this dataset, the commonly occurring nouns and adjectives referring to languages are often homonymous with nationality terms (e.g. *Deutsch* 'German', *Italienisch* 'Italian'). We want to be able to distinguish these uses.

Times & events This group covers events and temporal expressions. Motivated by the subject area of our oral narrative corpus, we add AGE as a special category distinct from cardinal numbers. We also note that in the narrative interview dataset, we find date and age specifications that are much less common in written texts or which represent the Austrian variety of German. Among them are abbreviated references to years by the last two digits as in 1, the Austrian term *Jänner* for the month of January (where German speakers in Germany use *Januar*) (cf. 2), or the special year-referring construction shown in 3

(1) Von *33* bis *36* konnten sie noch Möbel hierher bringen

[7] In OntoNotes, the category has the acronym NORP.

...
'From 1933 till 1936 they were still able to bring furniture here ... '

(2) jetzt im *Jänner* sind es sechzig Jahre
'this January it'll be sixty years'

(3) Im *47er Jahr* sind die Unruhen hier ausgebrochen.
'In the year '47, unrest erupted here.'

Quantitative This group covers numbers and quantities. For the news tweet data, we added the SCORE category to capture mentions of sports scores. Note that we keep instances of RATE that relate to the distribution of events over intervals in the temporal category FREQ (cf. 4), although formally they are simply sub-cases of the general RATE construction (cf. 5).

(4) [Einmal im Monat FREQ] ist ein hebräischer Vortrag.
'Once a month there is a talk in Hebrew.'

(5) Man hat [vier Scheiben Brot im Tag RATE] bekommen.
'You got four slices of bread per day.'

We also use a category MON for amounts of money. In the Nosta-D scheme, currency terms like *Euro* are treated as MISC. In our scheme, they are part of MON labels.

Miscellaneous The label MISC is available for entities that are not subsumed by one of the other labels.

Finally, as a general policy, we apply labels to instances as appropriate to the context. That is, if a LOC or GPE name is used metaphorically such as Paris in 6, we do not label it as an instance of its regular class. Similarly, when GPE names are used metonymically to refer to sports teams, as in (7), we label them as ORG.

(6) [Bucharest GPE] is the Paris of the east.

(7) [France ORG] is playing [Germany ORG] tonite.

We do, however, follow common practice in not differentiating between LOC and GPE depending on the context of a reference. Thus, Berlin in (9) gets the label GPE even though there it is not used to talk about the German government or the city's own municipal government, as in (8).

(8) [Berlin GPE] announced new tax plans today.

(9) My brother lives near [Berlin GPE].

4.2. Data set statistics

Table 2 shows the unnormalized counts for each category in the two datasets. (Recall that the ISW corpus has almost three times more tokens than the news dataset.) It is notable that the classes are not very similarly distributed. The Spearman's correlation between the number of instances in each corpus is only 0.24. For instance, PER and ORG names are significantly less frequent in the interviews than the news data. Mentions of LAN(guage) names, TIMEs and DUR(ation)s, by contrast, are much more frequent in the interviews than the news. These differences mostly reflect the differences between the genres. For instance, the interviews focus on the biographies of the speakers, therefore mentions of person names are relatively less frequent since the speakers' narratives involve many self-references and references to relatives with kin terms. Likewise, narration brings with it more (vague) TIME expressions than specific DATEs. Similarly, the speakers often talk about larger phases and events in their lives, motivating more specifications of DUR(ation), whereas the news headlines are mostly focused on specific current events.

4.3. Agreement

28 transcripts of the Israel corpus were annotated by annotator A, 60 were labeled by annotator B, with an overlap of 5 transcripts. On these 5 transcripts we achieved kappa values in a range from 0.748 to 0.793. The 83 transcripts were then adjudicated and checked for consistency using DECCA (Dickinson and Meurers, 2003) by annotator B, who had also led the development of the annotation guidelines. The social media data was annotated by annotator B. To allow for agreement testing, annotator C labeled 3000-token samples from two different newspapers, Krone and Allgemeine Zeitung. Agreement on the news tweets was much higher, reaching a kappa value 0.927 on the second sample. While some of the higher agreement may be due to a different label distribution, it also seems that the Twitter data contain less variety and fewer difficult cases for several classes, among them, for instance, temporal expressions.

4.4. Comparison to coarse-grained scheme

In order to have a better sense how our fine-grained scheme compares to the classic 4-category scheme (PER, LOC, ORG, MISC), we re-annotated a subset of 5.431 tokens from the GermEval 2014 dataset's training data following our guidelines. Since our fine-grained scheme currently only provides one layer of annotation, we compare to the top-level NER layer of the GermEval data.

First , we see that the fine-grained scheme with its larger label set provides more labels (569 vs 355) and covers more tokens (922 vs 539) than the coarse-grained scheme. For the Person class, there is no conceptual difference between the coarse and the fine-grained scheme. We judged one instance to denote a creature rather than a person, resulting in a minimal difference. Coarse-grained ORGs correspond to fine-grained ORGs 80% of the time. However, some of the cases where our fine-grained annotation seemingly deviates result from what we consider erroneous labels in the coarse-grained data. In the sequence *Ausbildung in der Bundesrepbulik Deutschland* 'education in the Federal Republic of Germany', the country name *Bundesrepbulik Deutschland* is mis-labeled as an ORG in the GermEval data, whereas we used GPE. Other cases may be more debatable. For instance, we treated *Olympische Spiele* 'Olympic Games' or *AchemAsia*, the name of trade show, as events (EVT) rather than ORGs. Coarse-grained LOCs mainly correspond to GPEs (geopolitical entities) in the fine-grained annotation of the dataset. Facilities account for about 10% of coarse-grained LOCs. Only about a quarter of coarse LOCs correspond to fine-grained LOCs. Of the coarse grained OTH(er) class for miscellaneous items, most instances correspond to works of art, in particular book titles, on the fine-grained

count	coarse label	fine equivalents
105	LOC	62 GPE , 24 LOC, 11 FAC, 1 EVT, 7 larger span
25	LOCderiv	25 NRP
10	LOCpart	6 NONE, 1 EVT, 1 GPE, 2 larger span
55	ORG	44 ORG, 3 EVT, 3 PRODUCT, 1 GPE, 1 MED, 2 NRP, 1 TITLE
15	ORGpart	7 NONE, 3 ORG, 5 TITLE, 2 larger span
43	OTH	11 ART, 2 EVT, 8 MISC, 1 ORG, 7 PRODUCT, 1 URL, 13 larger span
1	OTHderiv	1 NRP
93	PER	92 PER, 1 CREAT
2	PERderiv	1 NRP, 1 ORG
6	PERpart	1 NRP, 2 ORG, 3 NONE
254	NONE	76 DATE, 28 TITLE, 23 DUR, 20 TIME, 20 CARDINAL, 14 ORDINAL, 10 MON, 9 QUANT, 8 PERC, 7 AGE , 5 ORG , 34 others

Table 3: Correspondence between coarse grained labels and fine-grained labels in re-annotated GermEval 2014 subset

scheme, with product names also a noticeable subset. Still, for the fine-grained scheme, there remain instances that stay in the MISC class as they cannot be assigned to a more specific category.

Of the coarse subclasses for words that are derived from NEs (e.g. LOCderiv), most instances correspond to the fine-grained NRP class, as expected. For the subclasses representing NEs that are parts of words due to compounding (e.g. ORGpart), many instances go unlabeled in the fine-grained scheme as we have no sub-word annotation mechanism. However, this doesn't affect all such instances because on our scheme such words may still be part of a larger NE. For instance, for the NP *VW-Aufsichtrat Christian Wulff* 'VW-board member Christian Wulff', the coarse-grained scheme annotates the first token as ORGpart, whereas we treat the token as an instance of TITLE, as shown in (10).

(10) Coarse: [VW-Aufsichtrat ORGpart] [Christian Wulff PER]
Fine: [VW-Aufsichtrat TITLE] [Christian Wulff PER]

Finally, we see that the fine-grained annotation on the GermEval dataset mostly adds labels for temporal categories (DATE, DUR, TIME) and numeric categories (ORDINAL, CARDINAL,MON, QUANT, PERC).

5. Experiments

To establish baseline scores for tagging performance using our scheme, we experimented with two systems that model the task of NER as a sequence labeling problem. The one for which we report results here is a neural sequence tagger based on Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Our second system uses the character-based contextual string embeddings, provided by the the flair library (Akbik et al., 2018; Akbik et al., 2019). Since the flair tagger was consistently outperformed by the BERT-based system, we do not report results for it here for lack of space.

Transformers, of which BERT is an example, have recently pushed the state of the art for many NLP applications by learning context-sensitive embeddings with different optimization strategies and then fine-tuning the pre-trained embeddings in a task-specific setup. BERT embeddings are usually trained on large amounts of data, incorporating word embeddings with positional information and self-attention. The representations are trained in two different task setups, i.e. by predicting masked words based on their left and right context and by classifying two sentences based on how probable it is that the second one immediately succeeds the first one in a text document. As a result, the learned embeddings encode information about the left and right context for each word which makes them superior to most previous representations.

	Acc.	**Acc.**	**Prec.**	**Rec.**	**F1**
ID	**(all)**	**(non-O)**			
1	98.19	86.18	85.88	83.25	84.55
2	98.23	86.94	85.58	83.63	84.60
3	98.16	86.22	85.13	83.06	84.08
4	98.20	86.57	85.24	82.79	83.99
5	98.17	86.10	85.99	83.17	84.56
avg.	**98.19**	**86.40**	**85.56**	**83.18**	**84.36**

Table 4: Results for NER sequence tagging with BERT on the German CoNLL-2003 data.

TAG	**Prec.**	**Rec.**	**F1**
LOC	84.3	83.7	84.0
MISC	75.5	75.8	75.7
ORG	80.6	71.5	75.8
PER	95.3	94.5	94.8

Table 5: Label-wise results on the German CoNLL-2003 data (averages over 5 runs).

Devlin et al. (2019) have proposed a BERT architecture for sequence tagging on the CoNLL-2003 NER shared task data (Sang and Meulder, 2003). The model uses the pre-trained BERT embeddings for initialization and then fine-tunes the representations by adding a simple classification layer on top of the pre-trained BERT model and jointly optimizing the model parameters on the downstream task. Each BERT model provides its own tokenization which splits longer words into sub-tokens. The sequence tagger uses only the first sub-token as the input to the classifier, which then predicts a label for each token. In our experiments, we use the HuggingFace transformers library (Wolf et al., 2019) that provides pre-trained transformer models for different languages and tasks. We use the pre-trained cased German BERT model (bert-base-german-cased).[8]

To get a sense of the performance of this system, we first applied it to the well-established coarse-grained CoNLL 2003 dataset. Table 4 shows results for each of 5 runs of the BERT-based system. The results are about 4 points lower than the F-score of 88.27 reported as state of the art for the flair library but higher than our own results for flair when trying to replicate the state of the art. Table 5 reports the re-

[8] The model has been trained on the latest German Wikipedia dump (6GB of raw txt files), the OpenLegalData dump (2.4 GB) and news articles (3.6 GB). For details see `https://deepset.ai/german-bert`.

	Acc.	Acc.	Prec.	Rec.	F1
	(all)	(non-O)			
1	97.92	85.00	86.04	85.94	85.99
2	97.90	84.91	86.11	86.08	86.09
3	97.91	84.76	86.35	85.96	86.16
4	97.81	84.59	85.23	85.67	85.45
5	97.92	85.41	85.51	86.12	85.81
avg.	**97.89**	**84.93**	**85.85**	**85.95**	**85.9**

Table 6: Results for NER sequence tagging with BERT on the Israel corpus (last row reports avg. over 5 runs).

sults per label, showing that the more frequent classes PER and LOC are easier to learn on the CoNLL dataset than the ORG and MISC classes.

5.1. Israel corpus

We now turn the BERT system to the Israel corpus. Table 6 shows results for training on the ISW corpus. The overall performance is better than for the coarse-grained labels in the CoNLL data. Table 7 shows the breakdown per class. Naturally, the larger classes tend to perform better, as indicated by a Spearman's correlation of 0.57 for the relationship between the number of instances in the training data and F1-score. But some categories over- or underperform relative to their number of instances, for diverse reasons. For instance, CARDINAL and ORDINAL do better than expected whereas ORG and AGE do worse. ORG, for instance, is confused with FAC, which is plausible given that certain organizations like schools are strongly associated with buildings. In 11, the snowball fight involves members of the school community rather than the buildings per se, which is why the instances of *Unterberger Gymnasium* is tagged as ORG in the gold standard.

(11) Also es waren (-) Schneeballschlachten zwischen unserem äh Gymnasium und der Unterberger Gymnasium . . .
So there were (-) snowball fights between our uhm high school and the Unterberg high school . . .

The LANGUAGE class seemingly punches above its weight, reaching an F1-score close to 96% that only slightly lags behind the score for the GPE class which has a bit more than twice as many instances. Closer inspection of the data shows that the tag ambiguity for words that are tagged with LANGUAGE at least once is about the same as the tag ambiguity for words tagged as GPE at least once. But while potential LANGUAGE words may not be less ambiguous than potential GPE words, it is notable that the data contain more instances per word for the former (12.6) than for the latter (9.4).

In the case of AGE, inspection of the confusion matrix and errors for the best BERT run shows that discrimination against other classes suffers, for instance, because of aborted or shortened DATE expressions that make these latter look more similar to typical two-digit AGE values.

(12) Mein Schwiegervater der hat äh also Herzog der hat neunzehn neunzehnvier glaube ich ist er gestorben
My father-in-law he has uhm so Herzog he has nineteen ninteen-four believe I is he died
'My father-in-law, he, uhm Herzog, he died in nineteen, nineteen o'four, I believe.'

TAG	Prec.	Rec.	F1	N_{train}
ADD	0	0	0	4
AGE	72.6	69.0	70.7	295
ART	41.2	20.9	27.7	46
CARDINAL	87.5	90.5	89.0	368
DATE	77.9	83.5	80.6	798
DUR	62.8	65.0	63.9	1026
EVT	81.8	83.3	82.3	115
FAC	72.5	91.0	80.7	191
FRAC	61.7	28.6	38.6	15
FREQ	73.1	73.1	73.1	372
GPE	97.0	96.3	96.7	2704
LAN	95.6	96.2	95.9	1298
LAW	29.3	24.0	26.0	10
LOC	63.2	44.0	51.7	126
MED	75.0	45.0	53.5	9
MISC	51.0	48.7	49.7	109
MON	35.0	46.7	38.7	46
NRP	93.8	91.9	92.8	1787
ORDINAL	82.3	83.3	82.7	300
ORG	58.1	70.6	63.7	248
PER	81.0	84.0	82.5	652
PERC	85.5	88.0	86.4	23
PRODUCT	0	0	0	12
PROJ	60.0	30.0	40.0	6
QUANT	70.4	54.5	61.1	21
RATE	15.0	12.0	13.1	10
SORD	57.2	95.6	71.5	207
TIME	92.2	88.4	90.2	2726
TITLE	69.8	66.1	67.8	78

Table 7: Label-wise results on the Israel corpus (averages over 5 BERT runs). N = number of instances in training data.

	Acc.	Acc.	Prec.	Rec.	F1
ID	(all)	(non-O)			
1	96.54	88.29	87.97	88.51	88.24
2	96.55	87.87	87.97	88.03	88.00
3	96.57	87.85	88.48	88.20	88.34
4	96.59	88.13	87.57	88.20	87.88
5	96.75	88.69	88.27	88.75	88.51
avg.	**96.60**	**88.17**	**88.05**	**88.34**	**88.19**

Table 8: Results for NER sequence tagging with BERT on the Twitter News data.

5.2. Results on news tweets

Table 8 shows the results for training and testing BERT on the news tweets and table 9 the class-wise performance. Scores are notably higher than on the biographic interviews.

5.3. Cross domain testing

So far we have used our two new datasets on their own, training and testing on them separately. Now we train on

TAG	Prec.	Rec.	F1	N_{train}
ADD	n/a	n/a	n/a	0
AGE	88.0	94.1	90.9	121
ART	53.8	50.4	51.9	116
CARDINAL	89.7	94.8	92.1	280
CREAT	0.0	0.0	0.0	8
DATE	86.3	90.6	88.4	483
DUR	76.1	76.5	76.3	159
EVT	57.0	58.0	57.4	192
FAC	72.5	72.9	72.7	217
FRAC	0.0	0.0	0.0	5
FREQ	57.3	48.9	52.4	33
GPE	93.5	94.1	93.8	1365
LAN	0.0	0.0	0.0	6
LAW	63.9	66.0	64.8	36
LOC	64.2	63.2	63.7	186
MED	73.3	20.0	30.3	17
MISC	31.1	31.0	30.8	104
MON	85.8	89.0	87.3	63
NRP	88.3	86.4	87.5	477
ORDINAL	95.4	86.7	90.8	86
ORG	85.0	83.2	84.1	1247
PER	91.3	90.0	90.6	1148
PERC	84.6	91.4	87.8	11
PRODUCT	64.0	72.0	67.6	76
PROJ	10.0	6.7	8.0	19
QUANT	47.1	68.6	55.7	91
RATE	0.0	0.0	0.0	7
SCORE	80.0	100.0	86.7	16
SORD	82.1	91.5	86.6	91
TIME	85.1	89.2	87.1	365
TITLE	70.6	76.9	73.6	0
URL	100.0	100.0	100.0	2949

Table 9: Label-wise results on the news data (averages over 5 BERT runs). N = number of instances in training data.

	Acc.	Acc.	Prec.	Rec.	F1
ID	(all)		(non-O)		
1	95.33	62.43	69.23	61.90	65.36
2	95.14	59.70	68.40	59.06	63.38
3	95.32	62.06	67.33	61.72	64.40
4	95.25	62.04	66.16	62.01	64.02
5	95.20	61.43	66.53	61.14	63.72
avg.	**95.25**	**61.53**	**67.53**	**61.17**	**64.18**

Table 10: Results for NER sequence tagging with BERT; train on news data, test on Israel data.

TAG	Prec.	Rec.	F1
ADD	0.00	0.00	0.00
AGE	37.68	29.21	32.91
ART	41.67	11.63	18.18
CARDINAL	73.30	90.42	80.97
DATE	51.95	58.82	55.17
DUR	54.74	46.44	50.25
EVT	55.26	42.86	48.28
FAC	45.28	28.57	35.04
FRAC	0.00	0.00	0.00
FREQ	38.21	34.31	36.15
GPE	93.81	90.26	92.00
LAN	100.00	1.30	2.57
LAW	0.00	0.00	0.00
LOC	22.89	47.50	30.89
MED	0.00	0.00	0.00
MISC	8.33	3.33	4.76
MON	60.00	100.00	75.00
NRP	70.91	76.27	73.49
ORDINAL	62.79	75.00	68.35
ORG	42.50	64.56	51.26
PER	71.79	80.00	75.68
PERC	62.50	50.00	55.56
PRODUCT	33.33	100.00	50.00
PROJ	0.00	0.00	0.00
QUANT	75.00	81.82	78.26
RATE	0.00	0.00	0.00
SCORE	0.00	0.00	0.00
SORD	32.08	53.12	40.00
TIME	75.62	62.36	68.35
TITLE	30.00	13.04	18.18

Table 11: Label-wise results; train on news data , test on Israel data (averages over 5 BERT runs).

one and test on the other. Table 10 shows results for training on the news tweets and then testing on the ISW corpus. Results are significanlty worse.

Table 11 breaks down the performance by tag. When we compare the scores in Table 11 to those in Table 7, we see that for the GPE and CARDINAL categories, we can get acceptable performance. For LAN(guage) performance collapses, as the news tweets contain hardly any training instances. But even on categories with many instances, such as PER and ORG, we observe a notable drop. Classes with fewer instances are more sensitive to the domain difference. We do worse, for instance, on AGE, which is harder to learn on the ISW data than on the news data to begin with. Here, the instances in the biographic interviews can look very differently from what is found in the news tweets. Similiarly, while DATE performs at over 80% F1 in the in-domain setting, it suffers significantly in the cross-domain setting. Again this stems from differences between the domains. For instance, while years are written as numbers in the news data, they are quite often spelled out in words in the transcripts. Problems also arise from two-digit year mentions in the ISW transcripts such as *19* rather than *1919*. Tables 12 and 13 show results for the inverse experiment, training on the ISW corpus and then testing on the news tweets. Results are even worse. For many labels such as PRODUCT, the Israel data provide significantly fewer training instances than are found in the news tweets. In other instances, we observe the effect of domain differences. For instance, while the Israel data contains a lot of GPEs in the form of city and country names, a large set of them are distinct from the ones mentioned in the news tweets. Further, mentions in the news tweets may appear as hashtags (#Mannheim), and/or lowercased (#oberpfalz), which is not the case for the transcripts. Interestingly,

ID	Acc. (all)	Acc.	Prec.	Rec.	F1
		(non-O)			
1	85.16	37.97	66.48	32.98	44.09
2	85.28	38.67	65.75	33.70	44.56
3	85.29	38.59	67.47	33.70	44.95
4	85.39	39.93	65.99	34.94	45.69
5	85.30	38.70	67.26	33.67	44.87
avg.	**85.28**	**38.77**	**66.59**	**33.80**	**44.83**

Table 12: Results for NER sequence tagging with BERT; train on Israel data, test on news data

TAG	Prec.	Rec.	F1
ADD	0.00	0.00	0.00
AGE	73.57	72.35	72.83
ART	2.86	0.74	1.18
CARDINAL	77.25	90.24	83.2
CREAT	0	0	0
DATE	73.62	68.39	70.87
DUR	49.53	59.42	54.02
EVT	56.55	8.17	14.1
FAC	33.33	24.12	27.82
FRAC	0	0	0
FREQ	49.42	86.67	62.72
GPE	73.68	44.14	55.2
LAN	21.86	100	35.67
LAW	0	0	0
LOC	33.01	14.64	20.17
MED	0	0	0
MISC	19.68	6.45	9.37
MON	73.99	70	71.92
NRP	59.07	64.69	61.71
ORDINAL	85.03	80.61	82.66
ORG	60.41	22.88	33.14
PER	77.3	61.42	68.38
PERC	76.19	80	77.86
PRODUCT	2.22	1	1.38
PROJ	0	0	0
QUANT	43.51	65.71	51.68
RATE	0	0	0
SCORE	0	0	0
SORD	83.18	74.62	78.52
TIME	69.26	76.46	72.63
TITLE	85.19	9.06	16.1
URL	0	0	0

Table 13: Label-wise results; train on Israel data , test on news data (averages over 5 BERT runs).

for the AGE category, the drop in performance is much less pronounced when going in the direction from the biographic interviews to the news tweets than what we saw for the opposite direction. This makes sense as spoken language has more variety of AGE expressions than what occurs in news but covers a substantial set of what is found there.

6. Conclusion

We have introduced two new datasets for German NER. They are the first ones to use a fine-grained label inventory with 30 classes and cover two quite distinct domains, spoken language in the form of biographic interviews and news-related tweets. In an empirical comparison of parallel coarse-grained and fine-grained annotations on the GermEval 2014 dataset , we saw that the proposed fine-grained scheme is compatible with the coarse-grained one, introducing new labels for sub-classes for some of the traditional coarse classes and especially breaking out new categories from the coarse-grained MISC class. A significant difference to the coarse-grained scheme is that we introduce new labels in the domain of temporal and numeric expressions. We also established some baseline results for labeling texts from our domains according to the new schema. These results showed that the news tweets overall are somewhat easier to predict than the interview data. Cross-domain experiments underscored the differences between the two domains. In future work, we want to use the data in a multitask setup to see if this allows for better results on both domains. We also want to label additional varieties of spoken interaction to see how well NER generalizes across speech.

7. Acknowledgements

We would like to thank Hanna Strub for her support in performing the annotations.

8. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, page 1638–1649, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, page 54–59, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Benikova, D., Biemann, C., Kisselew, M., and Padó, S. (2014a). GermEval 2014 Named Entity Recognition Shared Task: Companion Paper. In *Proceedings of the KONVENS GermEval workshop*, pages 104–112, Hildesheim, Germany.

Benikova, D., Biemann, C., and Reznicek, M. (2014b). NoSta-d named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Betten, A. and Du-Nour, M. (2004). *Wir sind die Letzten. Fragt uns aus.* Psychosozial Verlag.

Betten, A., Du-nour, M., Graßl, S., and Dannerer, M. (1995). *Sprachbewahrung nach der Emigration: das Deutsch der 20er Jahre in Israel.* M. Niemeyer.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceed-*

ings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, pages 1247–1250, New York, NY, USA. ACM.

Brambilla, M. and Flinz, C. (2019). Orte und entgegengesetzte emotionen (liebe und hass) in einem korpus biographischer interviews (emigrantendeutsch in israel – wiener in jerusalem). *Studi Germanici*, 15/16:165–187.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dickinson, M. and Meurers, W. D. (2003). Detecting errors in part-of-speech annotation. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, April. Association for Computational Linguistics.

Flinz, C., (2019). *Multiword Units and N-Grams Naming FEAR in the Israel-Corpus*, pages 86–98. 09.

Gillick, D., Lazic, N., Ganchev, K., Kirchner, J., and Huynh, D. (2014). Context-dependent fine-grained entity type tagging. *CoRR*, abs/1412.1820.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A., and Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 114–118, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India.

Leitner, E., Rehm, G., and Moreno-Schneider, J. (2019). Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287. Springer.

Leonardi, S. and Thüne, E.-M., (2011). *Wurzeln, Schnitte, Webemuster. Textuelles Emotionspotential von Erzählmetaphern am Beispiel von Anne Bettens Interviewkorpus Emigrantendeutsch in Israel*, page 229–246. 01.

Leonardi, S., Eva-Maria, T., and Betten, A. (2016). *Emotionsausdruck und Erzählstrategien in narrativen Interviews: Analysen zu Gesprächsaufnahmen mit jüdischen Emigranten*. Königshausen & Neumann.

Ling, X. and Weld, D. (2012). Fine-grained entity recognition.

Rosset, S., Grouin, C., and Zweigenbaum, P. (2011). Entités nommées structurées : guide d'annotation quaero. Technical report. autres.

Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147.

Schiersch, M., Mironova, V., Schmitt, M., Thomas, P., Gabryszak, A., and Hennig, L. (2018). A german corpus for fine-grained named entity recognition and relation extraction of traffic and industry events. In *Proceedings of the 11th International Conference on Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-18), 11th, May 7-12, Miyazaki, Japan*. European Language Resources Association.

Schmidt, T. (2014). The database for spoken german - dgd2. Proceedings of the ninth conference on international language resources and evaluation (LREC'14), pages 1451 – 1457, Reykjavik. European Language Resources Association (ELRA).

Sekine, S. and Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM.

Telljohann, H., Hinrichs, E., Kübler, S., Zinsmeister, H., and Beck., K. (2012). Stylebook for the tüingen treebank of written german (tüba-d/z). Technical report, Universität Tübingen, Seminar für Sprachwissenschaft.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans et al., editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

9. Language Resource References

Linguistic Data Consortium. (1994). *ECI Multilingual Text*. Linguistic Data Consortium, ISLRN 511-168-567-582-5.

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