The Extended DIRNDL Corpus as a Resource for Automatic Coreference and Bridging Resolution

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler and Katrin Schweitzer
University of Stuttgart, Institute for Natural Language Processing (IMS)
Pfaffenwaldring 5B, 70569 Stuttgart, Germany
{anders.bjoerkelund,kerstin.eckart,arndt.riester,nadja.schauffler,katrin.schweitzer}@ims.uni-stuttgart.de

Abstract

DIRNDL is a spoken and written corpus based on German radio news, which features coreference and information-status annotation (including bridging anaphora and their antecedents), as well as prosodic information. We have recently extended DIRNDL with a fine-grained two-dimensional information status labeling scheme. We have also applied a state-of-the-art part-of-speech and morphology tagger to the corpus, as well as highly accurate constituency and dependency parsers. In the light of this development we believe that DIRNDL is an interesting resource for NLP researchers working on automatic coreference and bridging resolution. In order to enable and promote usage of the data, we make it available for download in an accessible tabular format, compatible with the formats used in the CoNLL and SemEval shared tasks on automatic coreference resolution.

Keywords: anaphora, prosody, corpus annotation

1. Introduction

The Discourse Information Radio News Database for Linguistic analysis (DIRNDL) is a spoken corpus resource of German radio news (ca. 50.000 tokens, 3221 sentences). In its first release (Eckart et al., 2012), it was manually annotated for referential information status, i.e. the given-new classification of referring expressions (Riester et al. 2010) as well as prosodic GToBI(S) labels (Mayer, 1995). Constituent-structure annotations originated with the XLE parser (Crouch et al., 2011) and the LFG grammar by Rohrer and Forst (2006).

Aligning spoken language with its written transcript (or text with one of its read realizations) in a single resource is challenging for several reasons. Obviously, speech has a temporal determination which written language lacks. Punctuation marks (e.g. decimal points/commas) and compound words may receive different tokenisations in the different processing pipelines for written and spoken language, respectively. Moreover, speech, even the one by trained newsreaders, is seldom flawless and contains disfluencies and slips of the tongue, which are not contained in the written transcripts. (If they were, this would cause trouble to the parser.) In DIRNDL, these problems are tackled by integrating both sets of data with their different tokenisations in a PostgreSQL database and providing an alignment that uses multiple links (e.g. for accidental repetitions).

The database has proven to be a valuable resource for testing linguistic hypotheses at the interface between discourse, information structure, morpho-syntax and prosody. For instance, Riester and Piontek (submitted) extract all adjective-noun sequences from the corpus, together with their prosodic realization, in order to test whether NPs with accented adjectives necessitate the existence of contrastive alternatives. Augurzky et al. (submitted) investigate the influence of segmental clashes on the frequency of prosodic phrase breaks at the transition between two referring expressions, and between nominal heads and their embedded arguments. Rosenberg et al. (2012) and Soto et al. (2013) use DIRNDL for training an automatic prosodic labeler.

We have recently improved DIRNDL by revising existing annotations, and by adding new annotation layers, e.g. constituent trees and dependency trees from Björkelund et al. (2013), and named entities using Finkel et al. (2005) and Faruqui and Padó (2010). Based on this extension, we extracted DIRNDL$_{\text{anaphora}}$ as a resource for evaluation of automatic coreference and bridging resolvers. We exported the corpus in three tabular formats, two from recent shared tasks on automatic coreference resolution, i.e. the SemEval 2010 (Recasens et al., 2010) and CoNLL 2012 (Pradhan et al., 2012) shared tasks, and a third tabular format containing additional annotation layers which are not represented in the CoNLL or SemEval format, but might be useful for the resolution task, e.g. information status labels and pitch accents. In this paper we describe the exported annotation layers as well as the formats used. The export is freely available for download.[1]

2. Annotation layers

In this section we review the various layers of annotation in the DIRNDL corpus and the new export. Table 1 gives an overview of the annotation layers in the DIRNDL corpus as described by Eckart et al. (2012) and in the new DIRNDL$_{\text{anaphora}}$ release.

2.1. Pragmatic annotations

The corpus was previously annotated for *referential information status* following Riester et al. (2010). These annotations were replaced by two-dimensional information-status annotations following the *RefLex* scheme (Baumann and Riester, 2012), which distinguishes between *referential* and *lexical* information status. For example, (1) contains a referentially given (coreferential) phrase which comes with new lexical material (a so-called *epithet*). By contrast, in (2) there is a referentially new phrase which features lexically given material. For the referential as well as for the lexical level, the corpus contains links between the anaphor and its antecedent.

[1] http://www.ims.uni-stuttgart.de/data/dirndl

DIRNDL (Eckart et al., 2012)	**DIRNDL$_{\text{anaphora}}$**
Pragmatic annotations	
information status according to Riester et al. (2010)	information status according to RefLex scheme (Baumann and Riester, 2012)
Prosodic annotations	
GToBI(S) labels for pitch pitch accents and boundary tones	GToBI(S) labels for accents and boundary tones (revised manual annotations)
Morpho-syntactic annotations	
constituent trees by XLE parser with LFG grammar of Rohrer and Forst (2006)	i) lemmas predicted by the Mate lemmatizer (Bohnet, 2010) ii) part-of-speech tags and morphological tags by MarMoT (Mueller et al., 2013) iii) constituent trees and dependency trees from Björkelund et al. (2013)
Semantic annotations	
	named entities predicted by the Stanford named entity recognizer (Finkel et al., 2005)

Table 1: Overview of annotation layers in the first DIRNDL release and in DIRNDL$_{\text{anaphora}}$

(1) DIRNDL s1730, 2007-26-03, 17:00: Ahtisaari plädiert für eine Unabhängigkeit des Kosovo unter internationaler Aufsicht. Dies sei die einzige politische und wirtschaftliche Option für die Zukunft [der [serbischen]$_{\text{L-NEW}}$ [Provinz]$_{\text{L-NEW}}$]$_{\text{R-GIVEN}}$
Ahtisaari is making the case for an independence of Kosovo under international control. This would be the only political and economic option for the future of the Serbian province

(2) DIRNDL s206, 2007-25-03, 06:00: Ein Erdbeben der Stärke 7,2 hat Zentral-Japan erschüttert. Auch im Inselstaat Vanuatu im Südpazifik wurden [zwei [Beben]$_{\text{L-GIVEN}}$ der [Stärke]$_{\text{L-GIVEN}}$ 7,1 und 7,3]$_{\text{R-NEW}}$ registriert.
An earthquake measuring 7.2 has hit Central Japan. Also in the island state of Vanuatu in the Southern Pacific two quakes measuring 7.1 and 7.3 have been registered.

By *referential information status* we refer to the classical notion of information status discussed in the literature, e.g. Prince (1981), Nissim et al. (2004), Riester et al. (2010). A referring expression is R-GIVEN if it is a coreference anaphor. The label R-BRIDGING indicates that we are dealing with a bridging anaphor (Asher and Lascarides, 1998; Poesio and Vieira, 1998), i.e. a non-coreferring but nevertheless context-dependent expression, e.g. the European Union ... [the member states]$_{\text{R-BRIDGING}}$.
Lexical information status captures semantic relations (e.g. a noun, verb or adjective is L-GIVEN if it is identical, a synonym or a hypernym of word contained in the context). An overview of the basic labels is shown in Table 2. These labels also have subcategories, and we refer the reader to Baumann and Riester (2012) for further details. Inter-annotator agreement on radio news data was determined in Riester and Baumann (2013) at $\kappa = .75$ for the referential level, and $\kappa = .64$ for the lexical level. For both levels of information status, Baumann and Riester (2013) show that increased givenness on both levels leads to a lower accent rate and/or the use of perceptually less prominent (e.g. L*) accents. In particular, it is well-known that (non-contrastive) coreferential anaphors in English and German are often deaccented, a fact which is well-described in the literature (see e.g. Halliday (1967), Schwarzschild (1999), Umbach (2002), Büring (2007) and many others). It is therefore likely that information about pitch accents will be a useful feature in coreference resolution.

2.2. Prosodic annotations: GToBI(S) labels

DIRNDL comprises information about intonation, i.e. the way an utterance is organized tonally. A group of intonation models – autosegmental-metrical models (essentially all based on Pierrehumbert, 1980) – is well accepted and widely used when describing prosody. For a subset of DIRNDL (approximately 5hrs of speech), tonal events were annotated manually according to an autosegmental intonation model for German (GToBI(S), cf. Mayer, 1995). Tonal events are pitch accents, which mark some of the words in a phrase as being more prominent than others, and boundary tones, which mark the tonal phrasing of the utterance. Essentially, a tonal event can be described as a local maximum or minimum in the intonation contour. Therefore, GToBI(S) labels describe the pitch contour by means of two levels, low (L) and high (H) representing regions in the speaker's register. That is, H describes a high local target (a peak) and L indicates a low local target in the contour. For example, a rising accent is composed of a low target on the accented syllable followed by a rise of the contour on the post-accented syllable, and is therefore labelled as L*H. Analogously, H*L marks a falling accent.

The GToBI(S) inventory also includes labels for the boundaries of tonal phrase: intermediate phrases, which are minor tonal phrases, are marked with the label "–", intonation phrases, which correspond to major tonal phrases, are marked with the label "%". The latter can also be marked with a tone if the contour rises, or falls, respectively, at the end of the phrase (H% or L%). Table 3 gives an overview of the complete label set.

Pitch accents are annotated on the syllable level. To make the annotations available on the word level, in DIRNDL each accent was enclosed in two "|" symbols and if several accents occurred on one word token, they were concatenated in the order of appearance. For example if a token on the word level was accented with a rising accent followed by a falling one, it is represented as |L*H||H*L|

For DIRNDL$_{\text{anaphora}}$, the GToBI(S) annotations were checked for plausibility using pitch accent shape information as retrieved by a parametric intonation model (Möhler, 2001), and corrected if necessary.

Referential information status		Lexical information status	
Units: referring expressions (NP/DP, PP)		Units: nouns, verbs, adjectives, adverbs	
Label	Description	Label	Description
R-GIVEN	coreferential anaphor	L-GIVEN	word identity / synonym / hypernym / holonym / superset
R-BRIDGING	non-coreferential context-dependent expression	L-ACCESSIBLE	hyponym / meronym / subset / otherwise related
R-UNUSED	definite discourse-new context-free expression	L-NEW	unrelated expression (within current news item)
R-NEW	specific indefinite		
R-GENERIC	generic definite or indefinite		
OTHER	e.g. cataphors		

Table 2: Overview basic RefLex scheme

Pitch accents		Boundary tones	
L*H	rise	%	intonation phrase boundary
H*L	fall	H%	high end of intonation phrase
H*	high peak with potential late fall	L%	low end of intonation phrase
L*	low target with potential late rise	%H	high beginning of intonation phrase
L*HL	rise-fall	-	intermediate phrase boundary
HH*L	early peak		
H*M	stylised contour		
!	diacritic for tonal declination		
*?	marker for uncertain accent placement		
? diacritic for uncertainty			

Table 3: Overview of the GToBI(S) inventory

2.3. Automatic morpho-syntactic annotations

The DIRNDL corpus was originally parsed with the XLE parser (Crouch et al., 2011) and the LFG grammar by Rohrer and Forst (2006). XLE provides deep LFG constituent structure analyses[2] but unfortunately yields fragmented parses in a substantial number of cases.

In order to provide additional, more robust syntactic information, as well as more fine-grained morphological annotations, we apply several other automatic tools to the corpus. Specifically, we added the following annotations: (i) automatically predicted part-of-speech tags and morphological tags, predicted with MarMoT (Mueller et al., 2013), which has been shown to outperform other available part-of-speech and morphology taggers; (ii) predicted lemmas, using the lemmatizer of the Mate tools toolkit (Bohnet, 2010), a state-of-the-art statistical lemmatizer; (iii) automatically predicted constituent trees with the constituent parser from Björkelund et al. (2013); (iv) automatically predicted dependency trees with the dependency parser from Björkelund et al. (2013). The constituent and dependency parsers by Björkelund et al. (2013) have shown state-of-the-art performance and recently obtained the best results for German in the recent SPMRL 2013 Shared Task on parsing of morphologically rich languages (Seddah et al., 2013). In contrast to the LFG parser, which is rule-based and driven by a grammar, all these tools are data-driven. They were all trained on the TiGer treebank (Brants et al., 2002; Seeker and Kuhn, 2012) and therefore provide annotations adhering to the TiGer annotation scheme. Example constituent and dependency analyses of a fragment of (1) is shown in Figures 1 and 2, respectively.

Since we have no gold standard annotations for these layers we are unable to evaluate the accuracy of these tools on the DIRNDL data set, however we refer the reader to the papers of the respective tools for evaluations on other data sets.

2.4. Automatic named entity annotations

Named entities are closely related to coreferentiality. In the RefLex scheme, named entities typically receive an R-UNUSED label[3] on their first occurrence, and an R-GIVEN label on subsequent occurrences. We added named entities using the Stanford named entity recognizer (Finkel et al., 2005). Specifically, we used the German model created by Faruqui and Padó (2010), which, in addition to standard

[2] LFG F-structures are currently not integrated in the database.

[3] R-UNUSED entities may be subclassified as to whether the annotator decides them to be KNOWN or UNKNOWN.

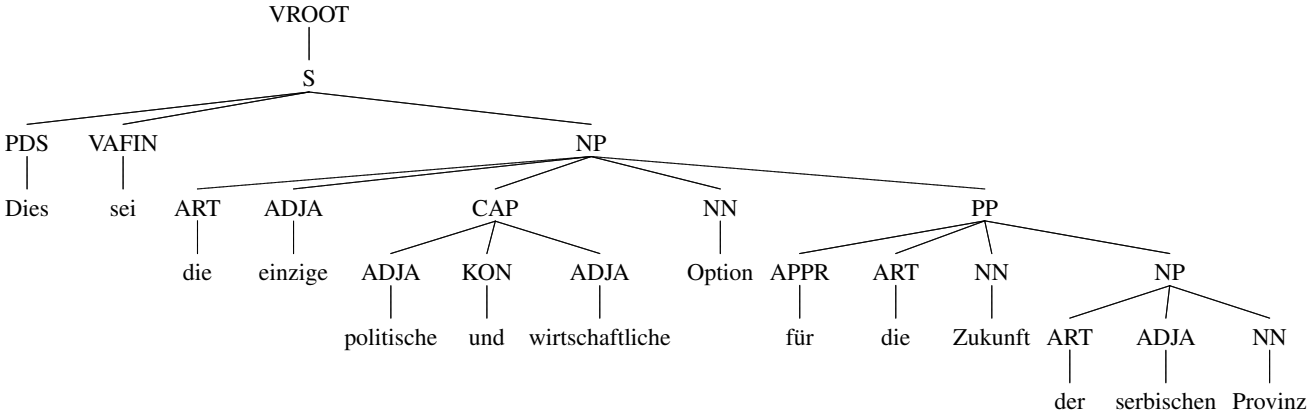

Figure 1: DIRNDL s1730, constituent tree

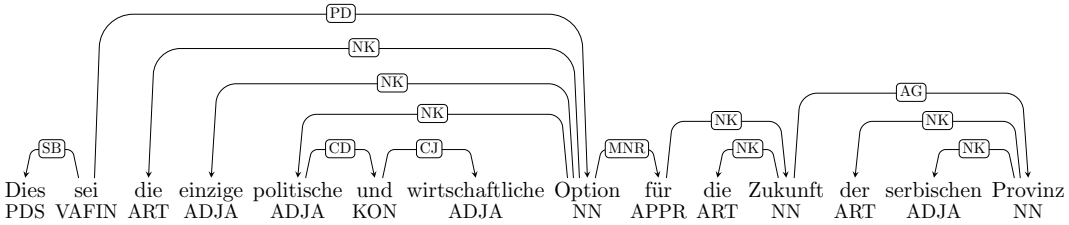

Figure 2: DIRNDL s1730, dependency tree

training data, also exploits large amounts of unlabeled data in its model.

3. DIRNDL export

This section describes the basic constitution of DIRNDL$_{\text{anaphora}}$ and gives an example of the tabular export format used in the release.

3.1. Constitution

The DIRNDL corpus consists of hourly radio news broadcasts from three days during 2007. The respective text transcripts were retrieved from the website of the corresponding radio station. The export of DIRNDL we describe in this paper does not contain the audio files of the spoken news, but is restricted to the transcripts.

It is important to note that, since the news broadcasts were consecutive, several items are repeated across broadcasts, sometimes with minor changes in between.[4] When using this resource either for training or testing automatic systems, we advise users to pay attention to these repetitions while conducting their experiments.

3.2. Tabular format

The original representation of the DIRNDL corpus is a relational database. While a relational database enables elaborate SQL queries, interfacing with a relational database is not the most convenient approach for NLP developers that are working on training and testing automatic systems. We therefore provide the new DIRNDL export in a tabular format, similar to the one used in the CoNLL 2011 and 2012 shared tasks (Pradhan et al., 2011; Pradhan et al., 2012). This also means that existing evaluation tools for automatic coreference can be used off the shelf against DIRNDL$_{\text{anaphora}}$.

An example of the tabular format is given in Figure 3, representing the two sentences from (1). The format represents each token on a single line and sentences separated by blank lines. Document boundaries are represented by the lines `#begin document` and `#end document`, where the former also contains a document identifier. In addition to the surface forms of each token, annotations are provided as additional columns in each line. A summary of the contents of the columns is displayed in Table 4. The first two columns hold document identifiers (document name and part); the following two hold sentence and token indexes, followed by the surface form of the word. Columns 6 through 9 hold predicted lemma, part-of-speech tag, morphological analysis, and named entity, respectively. The next three columns correspond to the syntactic structure: with the token index of the head word and the edge label according to the dependency tree (columns 10 and 11), followed by the constituent structure (column 12).

Columns 13 and 14 encode the prosodic features, as described in Section 2.2. i.e. the pitch accents, followed by the boundary tones. As outlined above, multiple pitch accents are concatenated. For instance, the second token in the first sentence (*UNO-Sondergesandte*) was realised with two pitch accents, a rising and a falling one, and was followed by an intermediate phrase boundary. If a word was realised without a pitch accent or without a boundary tone, the respective entry in the column is marked with the label "NONE". In the absence of an adequate mapping of the spoken realization and the textual tokenization, the label

[4] As part of the download package we provide a mapping from document identifiers to topics of the news items, which enables the extraction of repeated news items.

Column	Content	Manual/automatic
1	Document ID	-
2	Part number	-
3	Sentence no.	-
4	Token no.	-
5	Form	-
6	Lemma	A
7	Part-of-speech tag	A
8	Morphological features	A
9	Named Entity	A
10	Dependency head	A
11	Depdendency label	A
12	Constituent tree	A
13	Pitch accent	M
14	Boundary tone	M
15	Lexical IS	M
16	Referential IS	M
17	Coreference	M

Table 4: Column numbers, content our format

"N/A" was applied in $\text{DIRNDL}_{\text{anaphora}}$, for instance in the case of punctuation tokens or major deviations due to slips of the tongue. This label was also used for those cases, where no prosodic annotations were available.

The final three columns represent the pragmatic annotations: first the lexical layer (column 15), then the referential layer (column 16). The very last column encodes coreference, by grouping mentions into sets with common identifiers, as is the case in the CoNLL shared task format (Pradhan et al., 2012). For instance, the mention *des Kosovo* is R-UNUSED and belongs to the coreference cluster with id 901. The word *Kosovo* as such is labeled L-NEW. In addition to the referential and lexical information status labels, each mention that bears such a label has a unique identifier associated, separated by the `$` sign, i.e., *des Kosovo* has the identifier `6372` in the referential layer and *Kosovo* has the identifier `10513` in the lexical layer. The purpose of these identifiers is to simplify parsing the format in case of nested mentions, e.g. the full phrase *für eine Unabhängigkeit des Kosovo unter internationaler Aufsicht* is labeled R-BRIDGING-CONTAINED,[5] whereas the underlined subphrase is labeled R-GENERIC.

Since some of the RefLex labels have anchors in other mentions, these are also included as part of the labels in columns 15 and 16, separated by another `$`. For instance, the L-GIVEN-SUPER on *Provinz* in the second sentence indicates that this word is a hypernym of its anchor. The anchor is indicated by the last part of the label in column 15, `1-9-9`, which denote sentence number, first token, and last token, respectively. That is, the anchor for this label is the span *Kosovo* in the first sentence.

As mentioned above, in addition to the format described here, the release of the corpus also includes two other tabular versions of the same data. Specifically, it includes the tables in the CoNLL 2011/2012 shared task format as well as the SemEval 2012 shared task format.

[5]This label is a special case of R-BRIDGING, where the "antecedent" is contained within the referring expression itself.

4. Conclusion

We presented $\text{DIRNDL}_{\text{anaphora}}$, a resource for anaphora resolution created from the extended DIRNDL corpus, which contains spoken and written radio news amounting to roughly 50,000 words. The corpus has been manually annotated for prosodic and pragmatic information. The new version includes a revised and updated version of the pragmatic annotations, as well as automatic predictions by state-of-the-art morphosyntactic tools, including part-of-speech and morphology as well as dependency and phrase-structure syntactic trees.

Since our explicit goal is to enable developers of automatic tools for coreference and bridging resolution to use $\text{DIRNDL}_{\text{anaphora}}$ as a resource for evaluation, we are making the corpus available for download[6] in established text-based formats previously used for coreference resolution.

5. References

Asher, N. and Lascarides, A. (1998). Bridging. *Journal of Semantics*, 15:83–113.

Augurzky, P., Riester, A., and Tomaschek, F. (submitted). Segmental effects on prosody: modelling German argument structure. Phonetik & Phonologie 9, Zurich.

Baumann, S. and Riester, A. (2012). Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects. In Elordieta, G. and Prieto, P., editors, *Prosody and Meaning*, volume 25 of *Interface Explorations*, pages 119–162. Mouton de Gruyter, Berlin.

Baumann, S. and Riester, A. (2013). Coreference, Lexical Givenness and Prosody in German. *Lingua*, 136:16–37.

Björkelund, A., Cetinoglu, O., Farkas, R., Mueller, T., and Seeker, W. (2013). (re)ranking meets morphosyntax: State-of-the-art results from the SPMRL 2013 shared task. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 135–145, Seattle, Washington, USA, October. Association for Computational Linguistics.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In Hinrichs, E. and Simov, K., editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41, Sozopol, Bulgaria.

Büring, D. (2007). Intonation, semantics and information structure. In Ramchand, G. and Reiss, C., editors, *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press.

Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J., and Newman, P. (2011). XLE Documentation. http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html.

Eckart, K., Riester, A., and Schweitzer, K. (2012). A Discourse Information Radio News Database for Linguistic Analysis. In Chiarcos, C., Nordhoff, S., and Hellmann, S., editors, *Linked*

[6]`http://www.ims.uni-stuttgart.de/data/dirndl`

Data in Linguistics. Representing and Connecting Language Data and Language Metadata, pages 65–76. Springer, Heidelberg.

Faruqui, M. and Padó, S. (2010). Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Halliday, M. (1967). Notes on Transitivity and Theme in English. Part 2. *Journal of Linguistics*, 3:199–244.

Mayer, J. (1995). Transcription of German Intonation. The Stuttgart System. http://www.ims.uni-stuttgart.de/phonetik/joerg/labman/STGTsystem.html.

Möhler, G. (2001). Improvements of the PaIntE model for F_0 parametrization. Technical report, Institute of Natural Language Processing, University of Stuttgart. Draft version.

Mueller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October. Association for Computational Linguistics.

Nissim, M., Dingare, S., Carletta, J., and Steedman, M. (2004). An Annotation Scheme for Information Status in Dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisbon.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Ph.D. thesis, Massachusetts Institute of Technology, September.

Poesio, M. and Vieira, R. (1998). A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics*, 24(2):183–216.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, July. Association for Computational Linguistics.

Prince, E. F. (1981). Toward a Taxonomy of Given-New Information. In Cole, P., editor, *Radical Pragmatics*, pages 233–255. Academic Press, New York.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden, July. Association for Computational Linguistics.

Riester, A. and Baumann, S. (2013). Focus Triggers and Focus Types from a Corpus Perspective. *Dialogue & Discourse*, 4(2).

Riester, A. and Piontek, J. (submitted). Anarchy in the NP. When new nouns get deaccented and given nouns don't.

Riester, A., Lorenz, D., and Seemann, N. (2010). A Recursive Annotation Scheme for Referential Information Status. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 717–722, Valletta, Malta.

Rohrer, C. and Forst, M. (2006). Improving Coverage and Parsing Quality of a Large-Scale LFG for German. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genova.

Rosenberg, A., Cooper, E., Levitan, R., and Hirschberg, J. (2012). Cross-Language Prominence Detection. In *6th International Conference on Speech Prosody*, Shanghai.

Schwarzschild, R. (1999). GIVENness, AvoidF, and Other Constraints on the Placement of Accent. *Natural Language Semantics*, 7(2):141–177.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Marton, Y., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., and Wróblewska, A. (2013). Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA, October. Association for Computational Linguistics.

Seeker, W. and Kuhn, J. (2012). Making ellipses explicit in dependency conversion for a german treebank. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3132–3139, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1088.

Soto, V., Cooper, E., Rosenberg, A., and Hirschberg, J. (2013). Cross-Language Phrase Boundary Detection. In *Proceedings of ICASSP*, Vancouver.

Umbach, C. (2002). (De)accenting Definite Descriptions. *Theoretical Linguistics*, 27(2/3).

```
#begin document DIRNDL-200703261700 (part 000)
DIRNDL-200703261700  0  1   1  Der                 der                 ART    case=nom|number=sg|gender=masc           *       2  NK   (VROOT(S(NP*     |NONE|      NONE  -            (R-UNUSED-UNKNOWN$6371                    (899
DIRNDL-200703261700  0  1   2  UNO-Sondergesandte  uno-sondergesandt   NN     case=nom|number=sg|gender=masc           (MISC)  4  SB             *     |L*H||H*L|  -     (L-NEW$10510)  -                                         -
DIRNDL-200703261700  0  1   3  Ahtisaari           ahtisaari           NE     case=nom|number=sg|gender=masc           (PER)   2  NK             *)    |L*H|       -     (L-NEW$10509)  R-UNUSED-UNKNOWN$6371)                   899)
DIRNDL-200703261700  0  1   4  plädiert            plädieren           VVFIN  number=sg|person=3|tense=pres|mood=ind   *       0  --             *     |L*H|       NONE  (L-NEW$10514)  -                                         -
DIRNDL-200703261700  0  1   5  für                 für                 APPR   _                                        *       4  MO           (PP*    |NONE|      NONE  -            (R-BRIDGING-CONTAINED$6374                (900
DIRNDL-200703261700  0  1   6  eine                ein                 ART    case=acc|number=sg|gender=fem            *       7  NK             *     |NONE|      NONE  -            -                                         -
DIRNDL-200703261700  0  1   7  Unabhängigkeit      unabhängigkeit      NN     case=acc|number=sg|gender=fem            *       5  NK             *     |H*L|       NONE  (L-NEW$10512)  -                                         -
DIRNDL-200703261700  0  1   8  des                 der                 ART    case=gen|number=sg|gender=masc           *       9  NK           (NP*    |NONE|      NONE  -            (R-UNUSED-KNOWN$6372                      (901
DIRNDL-200703261700  0  1   9  Kosovo              kosovo              NE     case=gen|number=sg|gender=masc           (LOC)   7  AG             *)    |!H*L||L*H|  -     (L-NEW$10513)  R-UNUSED-KNOWN$6372)                      901)
DIRNDL-200703261700  0  1  10  unter               unter               APPR   _                                        *       7  MNR          (PP*    |NONE|      NONE  -            (R-GENERIC$6373                           -
DIRNDL-200703261700  0  1  11  internationaler      international       ADJA   case=dat|number=sg|gender=fem|degree=pos  *      12  NK             *     |H*L||L*H|  NONE  (L-NEW$10515)  -                                         -
DIRNDL-200703261700  0  1  12  Aufsicht            aufsicht            NN     case=dat|number=sg|gender=fem            *      10  NK             *)))  |!H*L|      %     (L-NEW$10511)  R-GENERIC$6373)|R-BRIDGING-CONTAINED$6374)  900)
DIRNDL-200703261700  0  1  13  .                   --                  $.     _                                        *       4  --             *)    N/A         N/A   -            -                                         -

DIRNDL-200703261700  0  2   1  Dies                dieser              PDS    case=nom|number=sg|gender=neut           *       2  SB   (VROOT(S*         |NONE|  NONE  -                        (R-GIVEN$6377)                                                  (900)
DIRNDL-200703261700  0  2   2  sei                 sein                VAFIN  number=sg|person=3|tense=pres|mood=subj   *      16  OC           *         |NONE|  NONE  -                        -                                                               -
DIRNDL-200703261700  0  2   3  die                 der                 ART    case=nom|number=sg|gender=fem            *       8  NK         (NP*        |NONE|  NONE  -                        (R-GENERIC$6382                                                 -
DIRNDL-200703261700  0  2   4  einzige             einzig              ADJA   case=nom|number=sg|gender=fem|degree=pos  *       8  NK           *         |L*H|   NONE  (L-NEW$10525)          -                                                               -
DIRNDL-200703261700  0  2   5  politische          politisch           ADJA   case=nom|number=sg|gender=fem|degree=pos  *       8  NK       (CAP*         |H*L|   NONE  (L-NEW$10527)          -                                                               -
DIRNDL-200703261700  0  2   6  und                 und                 KON    _                                        *       5  CD           *         |NONE|  NONE  -                        -                                                               -
DIRNDL-200703261700  0  2   7  wirtschaftliche     wirtschaftlich      ADJA   case=nom|number=sg|gender=fem|degree=pos  *       6  CJ           *)        |!H*L|  NONE  (L-NEW$10521)          -                                                               -
DIRNDL-200703261700  0  2   8  Option              option              NN     case=nom|number=sg|gender=fem            *       2  PD           *         |L*H|   -     (L-NEW$10523)          -                                                               -
DIRNDL-200703261700  0  2   9  für                 für                 APPR   _                                        *       8  MNR        (PP*        |NONE|  NONE  -                        (R-BRIDGING-CONTAINED$6376                                      -
DIRNDL-200703261700  0  2  10  die                 der                 ART    case=acc|number=sg|gender=fem            *      11  NK           *         |NONE|  NONE  -                        -                                                               -
DIRNDL-200703261700  0  2  11  Zukunft            zukunft             NN     case=acc|number=sg|gender=fem            *       9  NK           *         |H*L|   -     (L-NEW$10529)          -                                                               -
DIRNDL-200703261700  0  2  12  der                 der                 ART    case=gen|number=sg|gender=fem            *      14  NK         (NP*        |NONE|  NONE  -                        (R-GIVEN$6380                                                   (901
DIRNDL-200703261700  0  2  13  serbischen          serbisch            ADJA   case=gen|number=sg|gender=fem|degree=pos  (MISC)  14  NK           *         |H*L|   NONE  (L-NEW$10518)          -                                                               -
DIRNDL-200703261700  0  2  14  Provinz             provinz             NN     case=gen|number=sg|gender=fem            *      11  AG           *))))     |!H*L|  %     (L-GIVEN-SUPER$10516$1-9-9)  R-GIVEN$6380)|R-BRIDGING-CONTAINED$6376)|R-GENERIC$6382)  901)
DIRNDL-200703261700  0  2  15  ,                   --                  $,     _                                        *      16  --           *         N/A     N/A   -                        -                                                               -
DIRNDL-200703261700  0  2  16  heißt               heißen              VVFIN  number=sg|person=3|tense=pres|mood=ind   *       0  --         (S*          |L*H|   NONE  (L-NEW$10519)          -                                                               -
DIRNDL-200703261700  0  2  17  es                  es                  PPER   case=nom|number=sg|gender=neut|person=3   *      16  EP           *         |NONE|  NONE  -                        -                                                               -
DIRNDL-200703261700  0  2  18  in                  in                  APPR   _                                        *      16  MO         (PP*        |NONE|  NONE  -                        (R-BRIDGING-CONTAINED$6379                                      -
DIRNDL-200703261700  0  2  19  dem                 der                 ART    case=dat|number=sg|gender=masc           *      20  NK           *         |NONE|  NONE  -                        -                                                               -
DIRNDL-200703261700  0  2  20  Abschlussbericht    abschlussbericht    NN     case=dat|number=sg|gender=masc           *      18  NK           *         |L*H|   NONE  (L-NEW$10520)          -                                                               -
DIRNDL-200703261700  0  2  21  des                 der                 ART    case=gen|number=sg|gender=masc           *      24  NK         (NP*        |NONE|  NONE  -                        (R-GIVEN$6378                                                   (899
DIRNDL-200703261700  0  2  22  früheren            früh                ADJA   case=gen|number=sg|gender=masc|degree=comp *      24  NK           *         |H*L|   NONE  (L-NEW$10526)          -                                                               -
DIRNDL-200703261700  0  2  23  finnischen          finnisch            ADJA   case=acc|number=sg|gender=masc|degree=pos  (MISC)  24  NK           *         |NONE|  NONE  (L-NEW$10524)          -                                                               -
DIRNDL-200703261700  0  2  24  Präsidenten         präsident           NN     case=acc|number=sg|gender=masc           *      20  AG           *         |L*H|   -     (L-NEW$10528)          R-GIVEN$6378)                                                   899)
DIRNDL-200703261700  0  2  25  an                  an                  APPR   _                                        *      20  MNR        (PP*        |NONE|  NONE  -                        (R-BRIDGING-CONTAINED$6375                                      -
DIRNDL-200703261700  0  2  26  den                 der                 ART    case=acc|number=sg|gender=masc           *      27  NK           *         |NONE|  NONE  -                        -                                                               -
DIRNDL-200703261700  0  2  27  Sicherheitsrat      sicherheitsrat      NN     case=acc|number=sg|gender=masc           *      25  NK           *         |H*L|   NONE  (L-NEW$10522)          -                                                               -
DIRNDL-200703261700  0  2  28  der                 der                 ART    case=gen|number=pl|gender=fem            *      30  NK         (NP*        |NONE|  NONE  -                        (R-UNUSED-KNOWN$6381                                            -
DIRNDL-200703261700  0  2  29  Vereinten           vereint             ADJA   case=gen|number=pl|gender=fem|degree=pos  *      30  NK       (PN(NP*       |L*H|   NONE  (L-GIVEN-WHOLE$10517$2-27-27  -                                                               -
DIRNDL-200703261700  0  2  30  Nationen            nation              NN     case=gen|number=pl|gender=fem            *      27  AG           *)))))))  |!H*L|  %     L-GIVEN-WHOLE$10517$2-27-27)  R-UNUSED-KNOWN$6381)|R-BRIDGING-CONTAINED$6375)|R-BRIDGING-CONTAINED$6379)  -
DIRNDL-200703261700  0  2  31  .                   --                  $.     _                                        *      16  --           *)        N/A     N/A   -                        -                                                               -
```

Figure 3: Example of the export format