Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2401–2412
Marseille, 20-25 June 2022

Inducing Discourse Marker Inventories from Lexical Knowledge Graphs

Christian Chiarcos
Applied Computational Linguistics (ACoLi), Goethe University Frankfurt, Germany
Institute for Digital Humanities (IDH), University of Cologne, Germany
chiarcos@cs.uni-frankfurt.de

Abstract

Discourse marker inventories are lexical resources that define the meaning of discourse cues (discourse markers) in terms of associated discourse relation types. They are thus important tools for the development of both discourse parsers and corpora with discourse annotations. This paper explores the potential of massively multilingual lexical knowledge graphs to induce multilingual discourse marker lexicons by means of propagation methods. Given one or multiple source language discourse marker inventories and a large number of bilingual dictionaries to link them – directly or indirectly – with the target language, we study to what extent discourse marker induction can benefit from the integration of information from different sources, the impact of sense granularity and what limiting factors may need to be considered. Our study uses discourse marker inventories from nine European languages normalized against the discourse relation inventory of the Penn Discourse Treebank (PDTB), as well as three collections of machine-readable dictionaries with different characteristics, so that the interplay of a large number of factors can be studied.

Keywords: discourse marker, lexical knowledge graphs, lexical induction, OntoLex

1. Background

Discourse parsing has been a topic that received considerable renewed attention in the last years. The area does, however, still suffer from a general sparsity of resources. At the moment, we are aware of discourse corpora for little more than a dozen languages, only, including English (Carlson et al., 2003; Wolf and Gibson, 2005; Prasad et al., 2008), Chinese (Huang and Chen, 2011; Long et al., 2020), Czech (Poláková et al., 2013), Dutch (Van Der Vliet et al., 2011), French (Danlos et al., 2015), German (Stede, 2004; Gastel et al., 2011), Hindi (Oza et al., 2009), Italian (Raymond et al., 2007), Portuguese (Pardo and Nunes, 2004), Spanish (Da Cunha et al., 2011), and a small number of cross-linguistic discourse annotations, e.g. Buch-Kromann and Korzen (2010) and Zeyrek et al. (2019).

While for most languages, full-fledged discourse parsing is thus beyond reach, it is well-known that already more shallow techniques can be employed to facilitate the development of discourse-aware technologies. One such technique is the application of discourse marker inventories, i.e., machine-readable dictionaries of discourse cues ('markers') such as adverbs, conjunctions and particular phrases, classified according to fine-grained sense inventories that represent their discourse functions in terms of associated discourse relations. These can then be used as gazetteers to determine *possible* discourse functions of the associated marker in a text. A plain lookup may thus provide information about the discourse relation an utterance is subject to, or, at least, to narrow their possible bandwidth, and this information can either be directly used (as a replacement of discourse parsing) (Fuentes Fort, 2008) or used as a factor to support more sophisticated methods for discourse parsing (Bourgonje and Stede, 2020).

For bootstrapping discourse marker inventories, three main techniques have been applied so far: (1) extraction from a corpus with discourse annotation (Das et al., 2018), (2) induction from parallel corpora (Versley, 2010; Laali and Kosseim, 2014), or (3) translation of an existing discourse marker inventory in another language by the hands of a native speaker and/or existing dictionaries (Alonso, 2005).

In the creation of most discourse marker inventories we are aware of, these techniques have been applied in combination, but in most cases, the ultimate goal was the creation of a discourse-annotated corpus, not the creation of a discourse marker inventory per se, and thus, monolingual (1) or parallel (2) corpora have played a particular prominent role in this context, whereas lexical methods (3) have been somewhat neglected – at least as far as the *automated* induction of discourse marker inventories is considered. To some extent, this is motivated by the insight that general-purpose dictionaries can only cover *lexical* discourse markers, as phrasal discourse markers like *for this reason* would not normally be listed in a print dictionary. And, indeed, this is an inherent limitation. At the same time, a large number of discourse markers listed in the inventories we studied consisted of single words only, so that lexical induction methods are capable of providing at least a partial discourse marker inventory.[1] Al-

[1]Even within the same language, we find a great deal of variation in this regard, so that, for example, the PDTB v.2 inventory (Prasad et al., 2008; Chiarcos and Ionov, 2021) features 71% single-word expressions (54/186), whereas in the English DiscMar inventory (Alonso, 2005), only 45% of the listed discourse markers are single-word expressions (47/86). Overall, this seems to be largely a difference in lexical coverage, although theoretical considerations may play be a role,

ready a partial discourse marker inventory represents an invaluable resource for a language in which no discourse annotation exists, be it as a basis for the development of annotation guidelines (e.g., by giving a practical definition of discourse relations by means of a substitution test in the target language), or for evaluation and quality control for other methods (e.g., manual corpus annotation or projection in parallel corpora).

We address this gap and describe the application of large-scale lexical networks to automatize the translation-based method for discourse marker induction. We consider this particularly valuable for low resource languages where standard means of induction or projection via parallel corpora are either not possible (because of the lack of electronically available and legally cleared translated text) or restricted to highly specialized genres that are not representative for the language (e.g., Bible translations, subtitles, technical documentation). We follow the general setup of a series of Shared Tasks on Translation Inference Across Dictionaries (Ordan et al., 2017, TIAD) conducted continuously since 2017. The task is to take a collection of bilingual dictionaries (say, English-Spanish, English-Catalan, Spanish-French and French-Catalan) to bootstrap a new bilingual dictionary between languages indirectly connected by the resulting graph (say, English-Spanish) by using other languages (here, Spanish and Catalan) as pivots. Our scenario is structurally similar, as we propagate labels (rather than translations) from source language(s) to target language(s) by means of one or multiple pivot languages, so we adopt the technical setting of TIAD shared tasks, and in particular, approaches developed in this context that were based on the propagation of translations or concepts from source to target languages (Chiarcos et al., 2020b).

As we explore multi-source induction over multiple pivot languages, our method of builds heavily on the availability of two kinds of resources: (a) large-scale lexical data in consistent machine-readable formats, and (b) several discourse marker inventories using a single and consistent taxonomy of cross-linguistically equivalent discourse relations. Both kinds of data have previously been made available using compatible web standards, so that now, their conjoint evaluation becomes a relatively easy task.

2. Data

For both lexical data and discourse marker inventories, we operate on machine-readable editions on the basis of RDF (Klyne et al., 2004) and OntoLex-Lemon (Cimiano et al., 2016).

RDF, the Resource Description Framework,[2] is a W3C standard that provides a generic data model for directed labeled graphs on the web: Nodes, vertices and graphs are identified by Uniform Resource Identifiers (URIs, resp., Internationalized Resource Identifiers, IRIs), and on the basis of *HTTP-resolvable* URIs, a technical ecosystem evolved that facilitates the access to distributed data sets by standardized means of access (HTTP), query (SPARQL), and, in particular, the linking between datasets distributed on the web, hence the term 'Linked Data' (Berners-Lee, 2006). A notable feature of Linked Data technology is that the query language SPARQL can be used to query *across* different datasets, even if hosted by different providers. So, publishing data in RDF allows us to easily integrate information from different resources, and here, lexical data sets and discourse marker inventories.

Numerous RDF vocabularies define domain-specific data models, and OntoLex-Lemon[3] is a widely used community standard for publishing lexical resources as RDF data on the web (Cimiano et al., 2020, p.45-59). As such, OntoLex-Lemon has been applied in the aforementioned series of Shared Tasks on Translation Inference Across Dictionaries (TIAD). Both lexical data and discourse marker inventories are available in OntoLex editions.

As lexical data basis, we use the ACoLi Dictionary Graph (Chiarcos et al., 2020a), an aggregate over several major collections of bilingual dictionaries, e.g., Apertium (Forcada et al., 2011), FreeDict (Bański and Wójtowicz, 2009), the Open Multilingual WordNet (Bond and Foster, 2013), and PanLex (Westphal et al., 2015). Overall, it features more than 3,000 bilingual dictionaries from various sources, provided as RDF data in accordance with the Ontolex-Lemon vocabulary. We evaluate against three subsets from the ACoLi Dictionary Graph with specific characteristics:

Apertium (A) 53 bilingual dictionaries for 44 languages, high-quality datasets developed for machine translation.

FreeDict (F) 145 bilingual dictionaries for 104 language pairs and 45 languages.

MUSE (M) 108 dictionaries for 57 language pairs for 45 languages, predominantly linked via English.

These collections provide bilingual dictionaries for 174 language pairs and 77 languages (see Appendix A). For evaluation, we did not experiment with PanLex dictionaries (that constitute about 2/3 of the ACoLi Dictionary Graph) because these represent a rather uneven pool of language resources with heterogeneous characteristics and varying levels of quality. The subsets above, on the other hand, represent prototypical categories of dictionaries which allows us to explore the impact of their respective characteristics. They are, however, not fully comparable as they cover different languages.

e.g., in what consitutes a discourse marker and what constitutes an 'alternative lexicalization' of the respective discourse relation.

[2] https://www.w3.org/RDF/

[3] https://www.w3.org/2016/05/ontolex/

The second major component is a collection of discourse marker inventories that serve as a basis for subsequent induction. Here, we operate with discourse marker inventories for nine languages, DimLex (Stede and Umbach, 1998; Scheffler and Stede, 2016, German), DiscMar (Alonso, 2005, Catalan, English, Spanish), DisCo (Bourgonje et al., 2018, Dutch), LDM-PT (Mendes et al., 2018, Portuguese), LexConn (Roze et al., 2012, French), LICO (Feltracco et al., 2016, Italian), PDTB (Prasad et al., 2008, English), and PDiTB (Zikánová et al., 2019, Czech). Based on efforts conducted in the context of the TextLink network (Degand, 2016) to develop a unified XML representation for the majority of these inventories, Chiarcos and Ionov (2021) provide a Linked Open Data edition of this data, also on the basis of the OntoLex vocabulary. A key feature of this edition is that all inventories are linked with ontologies that define their sense inventories. Most inventories provide sense definitions modelled after the Penn Discourse Treebank (PDTB v.2 or v.3), but some follow independent conventions (DiscMar, LexConn, PDiTB). These ontologies are part of a larger collection of annotation models (ontologies) for PDTB and other forms of discourse annotation, all linked with a more general domain ontology for discourse annotations (Chiarcos, 2014). Via this domain ontology, it is then possible to derive an automated mapping between all supported schemas.

The PDTB taxonomy defines three primary levels of granularity as illustrated in Fig. 1. To these we add level 0 to express whether an expression can represent a discourse marker. Where an inventory does not provide a fine-grained distinction (as *weiterhin* in Fig. 1), coarse-grained labels from higher levels are being used. Note that not all inventories provide the full depth of the PDTB taxonomy, e.g., the DiscMar taxonomy roughly corresponds to the PDTB level 1 classification.

From the original OntoLex data, we use SPARQL queries to create tabular data with tab-separated values for both lexical data and discourse marker inventories. The query for lexical data retrieves a table of source language expression and target language expressions, generalizing over different modelling options in OntoLex-Lemon. The language of the expressions is encoded in BCP47 language tags. The query for discourse marker inventories retrieves data as structured in Fig. 1, i.e., source language expression and PDTB level classification. It is to be noted that the query consults the OntoLex-edited data which provides a link with the annotation model that defines the underlying taxonomy. If this is not the PDTB v.2 ontology, but another ontology, say, for the Rhetorical Structure Theory (Mann, William C. and Thompson, Sandra A., 1986, RST), the SPARQL query retrieves the linking of the RST model to the overarching reference model, and, then, indirectly, to the PDTB ontology. For every original RST concept, the mapping returns the PDTB concept(s) with the shortest path.

3. Induction by Sense Propagation

For discourse marker induction, discourse marker inventories and bi-dictionaries are connected into a single graph, which we iterate over different target languages. We operate under the assumption that the discourse marker inventories are exhaustive. That is, every lexical entry that does not conform to a known discourse marker is considered to not be a discourse marker. This is represented as a fifth top-level category in the sense hierarchy, at the same level as PDTB *EXPANSION*, *CONTIGENCY*, *TEMPORAL* and *COMPARISON*. Although this is an idealistic assumption, all discourse marker inventories considered here have been designed with the goal to provide such an exhaustive list, either in terms of lexical coverage or by being defined against a representative corpus.
The induction of discourse relations is calculated independently over all four levels. Prior to induction, we calculate the symmetric closure of dictionaries and merge all dictionaries that connect the same pair of languages l, k into a single dictionary $D_{l,k}$ (= $D_{k,l}$).

1. Be t the target language. Let the mapping function $M : \Sigma^+ \mapsto (R \mapsto [0:1])$ map any word w (in any language) to a mapping of PDTB relations R to a numerical score. Initially, M is empty. Be M^t the subset of M that applies to words from language t, the goal is to induce a mapping M^t that assigns the majority of words from t a sense mapping. If no such mapping can be found for any word $w \in t$ after induction, assume that $M^t(w) = (\textbf{true} \mapsto 0.0)$.[4]

2. For every source language s with a discourse marker inventory I^s, initialize the discourse mapping m^s as follows:

   $m^s(w^s) =$

   $$\begin{cases} \textbf{true} \mapsto 0.0 & \textbf{if } w^s \text{not in} I^s \\ rel \mapsto 1/length(I^s(w^s)) & \textbf{for all } rel \in I^s(w^s) \end{cases}$$

   Words from the dictionary that are not covered by the discourse marker inventory are initialized with a zero score for the sense **true**. Words or expressions covered by the discourse marker inventory are initialized with equal probability for all attested senses. If multiple discourse relations are given for a word w, the score is divided by their number ($length(I^s(w^s))$), so that unambigous discourse markers yield higher scores for their information. Add the mapping $m^s(w^s)$ to M and iterate for all attested words of all source languages.

[4]Note that every word w is also *typed* for its language in accordance with Turtle conventions, i.e., `"tree"@en` for the string *tree* in English. So, homographs from different languages are not being conflated by M.

```
# FORM          level 0 level 1     level 2                level 3
"weil"@de       True    CONTINGENCY CONTINGENCY:Cause      CONTINGENCY:Cause:Reason
"weiterhin"@de True    EXPANSION   EXPANSION:Conjunction EXPANSION:Conjunction
```

Figure 1: German discourse markers with PDTB senses at three levels of granularity

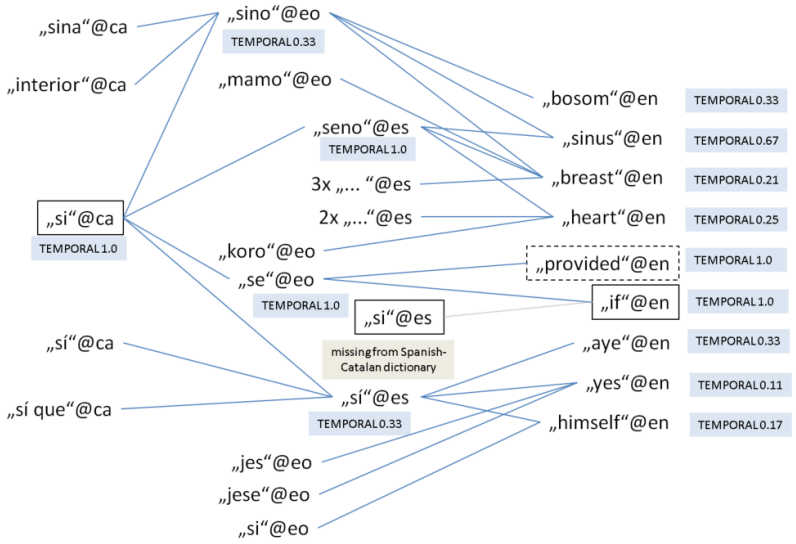

Figure 2: Induction of PDTB level 1 sense *TEMPORAL* from Catalan (DiscMar) to English via Esperanto and Spanish, Apertium dictionaries.

src	dict	#	p	r	f
level 0					
es	A	117	.47	.82	.60
es	F	87	.25	.73	.38
es	M	22	.59	.93	.72
es	all	169	.33	.84	.48
ca	A	138	.40	.78	.53
ca	M	14	.79	.85	.82
ca	all	140	.39	.78	.52
level 1					
es	A	131	.44	.82	.57
es	F	108	.20	.73	.32
es	M	27	.52	.93	.67
es	all	203	.29	.83	.43
ca	A	155	.37	.76	.50
ca	M	16	.81	.87	.84
ca	all	157	.37	.76	.50

Table 1: Direct induction experiments for the English DiscMar inventory, PDTB levels 0 and 1 for A(pertium), F(reeDict), M(USE) and all dictionaries)

3. For every language l and every attested word w^l from that language, be $D(w^l)$ the set of translations provided for the word w^l in any existing bi-dictionary. If $M(w^l)$ is undefined, then: For every word $v \in D(w^l)$ and $M(v)$, the set of mappings from v to PDTB senses and their scores; if there is any $v \in D(w^l)$ for which $M(v)$ is defined, set

$$m(w^l) = rel \mapsto \sum_{v \in D(w^l)} \begin{cases} 0 \textbf{ if } rel \notin domain(M(v)) \\ \frac{M(v)(rel)}{length(D(w^l))} \textbf{ otherwise} \end{cases}$$

So, the sense assignments and their scores are averaged over all translations of w^l for which a sense mapping has been previously established.

Add $m(w^l)$ to M and iterate over all attested words of all languages in the bi-dictionaries until no further additions to M are possible.

4. Return M^t as the resulting mapping from target language word forms to PDTB discourse relations and their respective senses.

The resulting sense mapping is reduced to word forms for which one or more discourse senses receives a score greater than 0. Figure 2 illustrates the procedure for a Catalan-English example: Black boxes indicate discourse markers according to DiscMar, the dashed box indicates a partial overlap with a discourse marker. (Note that errors in the dictionaries, e.g., the typo *sí* for Spanish *si*, have not been fixed for the figure.)

This algorithm is refined by constraints as described below. We evaluate precision, recall and f measure relative to the number of senses predicted or missed. The evaluation at level 0 is equivalent to token- (rather than sense-)level precision and recall. Recall is calculated as the number of correctly predicted senses relative to the number of senses of all target language words *in the dictionary*. By excluding out-of-vocabulary words, we limit the evaluation to *lexical discourse markers*. Phrasal discourse markers play an important role in communication, but are inaccessible by this method and thus excluded from the evaluation.

4. Evaluation

We evaluated all combinations of source languages, target languages, dictionary sets and parameters for all four PDTB levels for direct and indirect induction, more than 250,000 combinations in total. Only selected evaluation results can be included here, but result tables are chosen to be representative for all combinations of source and target languages.

As we aim to assess the multilingual dimension of discourse marker induction, direct induction (i.e., translation of discourse markers without using pivot languages) allows us to estimate *upper* quality bounds for what can be expected from indirect translations over multiple or an unconstrained number of pivot languages.

4.1. Direct Induction

In the direct induction scenario, we take one source language discourse marker inventory and perform a direct, lookup-based translation into the target language, without iterating over pivot languages or including multiple source languages. This naive setting and presumable gold standard is the conventional way for creating a discourse marker inventory from scratch, although normally, in a manual process.

Table 1 summarizes example results for direct induction of an English discourse marker inventory from Spanish, resp., Catalan, restricted to PDTB levels 0 and 1. Here, we operate with the DiscMar inventory, as by having the same author, this represents a particularly consistent subset of discourse marker inventories and the languages it covers allow to compare all three dictionary families directly. Comparable results (but over multiple source inventories) for PDTB-based discourse marker inventories are shown in Tab. 3, A main insight here is that the characteristics of the bi-dictionaries are essential for induction quality. At the same time, coverage of the individual dictionary collections is limited, to that subsequent experiments are primarily evaluated for the all-dict setting.

In terms of precision and recall, we observe similar results for discourse marker induction using Apertium for both source languages. The comparative decay of FreeDict reflects the limited coverage of the English-Spanish FreeDict dictionaries – which feature only less than 9,000 translation pairs per direction. As for MUSE, the relatively good performance in precision and recall is clearly misleading (as evident from the small number of total predictions), and can be attributed to the nature of alignment-induced word lists, as these tend to produce 1:1 correspondents, MUSE has a strong bias against phrases. Another factor is that this is a *high-precision* excerpt from a translation table, i.e., entries with low precision have been largely filtered away. For discourse markers, whose meaning is to a large extent contextually determined, high-precision alignment is harder to achieve than for designations of concepts (nouns), states (adjectives) and events (verbs). As a result, only close correspondents of single-word discourse markers are being induced. As these are also captured by more full-fledged lexical resources such as Apertium, adding MUSE to Apertium has no positive effect (as shown for induction from Catalan). Indeed, the level of noise introduced by MUSE seems to have a negative effect (as shown for induction from Spanish in the all-dict setting on comparison to the Apertium setting).

In the full PDTB setting, no Apertium dictionary is available that permits to directly translate any native PDTB inventory into another language. We thus evaluate FreeDict and MUSE only. The highest f scores were 0.66 for level 0 (Dutch-English), 0.60 for level 1 (Italian-German), 0.54 for level 2 (German-English) and 0.48 for level 3 (Italian-German), all obtained by MUSE, but (as in the DiscMar experiment) with insufficient coverage in terms of predicted discourse markers. That these numbers are substantially lower than the scores obtained for DiscMar probably reflects structural and conceptional differences in the inventories that their mapping to PDTB relations could not fully compensate.

4.2. Refinements, Filters and Parameters

In comparison to direct induction from a single source inventory, which is available under certain circumstances only, indirect induction over one or multiple pivot languages allows to substantially extend the range of languages for which discourse marker inventories can be induced. Moreover, it allows to more flexibly support induction from multiple source inventories in different languages, which may help to disambiguate and refine each others predictions. We assume that (1) induction over one or multiple pivot languages allows to circumvent sparsity issues, and that (2) conjoint induction from multiple inventories (via direct or indirect translations) increases prediction quality. However, induction over longer sequences of dictionaries are also more error-prone, as lexical ambiguity will naturally lead to a decay in confidence. For further improvement, we thus introduce a number of parameters for constrained induction:

min score after induction, return only senses (discourse relations) with a score greater than a predefined threshold τ

min pivots if during induction, a given word w does not have a sense assignment yet, then infer a sense assignment if and only if translations in at least m translations exist for which a sense mapping is defined in M.

min pivot languages for a given word w, require that translations in at least n languages exist for which a sense mapping is defined in M. The difference to the min pivots restriction (which operates on pivot words, not languages) is that a language restriction requires *independent evidence* from different lexical resources.

4.3. Induction over the Full Lexical Graph

For induction from multiple sources over multiple dictionaries, we assume that prediction quality increases with the number of parallel translations involved. However, feeding in additional discourse marker lexicons (and thus, more translation paths) does not immediately improve the picture. Evaluation was conducted over the respective full dictionary inventory and is reported here for the induction from all discourse marker inventories (except the English PDTB inventory) to the English DiscMar inventory.

Table 2 summarizes the evaluation results with the best-performing configuration (in terms of f measure) per

feature. In unconstrained induction, without restrictions on pivots, all induction results have much lower f scores than direct induction, but when induction is constrained, results reach or even exceeed the best case performance obtained for direct translation.

As for min pivots, we find that f scores reach or exceed the best case performance direct induction/translation in the all-dict setting. Although the source data is different (and more heterogeneous) than that taken as basis for Tab. 1, the evaluation basis is the same, so they can be directly compared. For Apertium and MUSE, the best results were achieved with a min pivot restriction of 2, resp., 3; for FreeDict, only with min pivot restriction of 10 (the maximum tested). We attribute this to differences in coverage of many FreeDict dictionaries which can be partially compensated by relying on longer induction paths (as resulting from high min pivot restrictions): The median size of FreeDict dictionaries is 15,537 translation pairs, but the actual size varies between 140 and 671,447 translation pairs, so that many smaller languages are not effectively covered, even though they are in the dictionary graph. For Apertium, median size is 22,285 translation pairs, but much more balanced, ranging from 4,693 to 106,880 translation pairs. For MUSE, this is similar, with median 17,133 translation pairs, ranging from 4,082 to 91,849 translation pairs.

As for min pivot languages, we, again, tested for thresholds from 1 to 10. For Apertium and FreeDict, we confirm a comparable increase in f score. For MUSE, the pivot language restrictions had no positive effect. We consider this to be an artifact of the characteristics of MUSE, which provides direct links with English for all languages and has a sampling bias against ambiguous discourse markers (as part of the automated pruning applied).

Finally, both conditions have been combined on the all dictionary setting. Here, min pivot languages was found to be the determining factor, with additional min pivot restrictions not leading to any improvements. This is consistent with expectations about the structure of the dictionary graph: High min pivot restrictions posit positive conditions for *either* evidence from additional sources, *or* ambiguous translation pairs. Increasing min pivot restrictions beyond the number of available incoming dictionaries will thus prevent unambiguous translation pairs from contributing to the induction. In our data, this effect outweighs any possible effects to be expected from excluding small-scale dictionaries that create a 'shortcut' between languages connected by more large-scale dictionaries with an indirect path. In subsequent experiments, we thus optimized the pivot language threshold, but not the min pivots threshold.

The improvements over direct induction are most significant on the Apertium data, less so on MUSE, but neither prominently in FreeDict nor in the all dictionary setting. Constrained induction in the all dictionaries setting fails to achieve the performance of the Apertium data set that it contains. In all configurations tested so far, its f scores are *below* those of any specific subset of dictionaries here, and we take this to be indicative of a deeper problem with differences in coverage and structure among the bilingual dictionaries considered.

dict	min score	min pivots	min pivot languages	#	p	r	f
no pivot restrictions							
A	0.15	1	1	230	.31	.88	.46
F	0.35	1	1	1329	.03	.49	.06
M	0.45	1	1	79	.24	.87	.38
all	0.4	1	1	1111	.04	.55	.08
best-performing min pivots, from 1 to 10							
A	0.2	2	1	164	.38	.86	.53
F	0.4	10	1	196	.13	.47	.20
M	0.45	2	1	77	.25	.91	.39
all	0.45	10	1	186	.15	.48	.23
best-performing min pivot languages, from 1 to 10							
A	0.15	1	2	159	.42	.93	.57
F	0.4	1	7	107	.17	.58	.26
M	0.45	1	1	79	.24	.86	.38
all	0.4	1	4	290	.14	.58	.22
best-performing combination							
all	0.4	4	4	290	.14	.58	.22

Table 2: Induction from all discourse marker inventories (except English) to English DiscMar, level 1

4.4. Level 2 and 3 Senses

For projecting the senses of deeper levels of the PDTB taxonomy, we exclude the DiscMar inventories (that do not provide this information to the full extent) and report results for multi-source induction over the full dictionary graphs for German DimLex and English PDTB inventories. These represent the two major types of discourse marker inventories, with DimLex representing the lexicographical tradition (albeit enriched with corpus data) and PDTB representing corpus-based discourse marker inventories (albeit informed by lexicographic research).

We report the best-performing configurations for the parameters introduced so far. In addition, we now normally project more than one sense per discourse marker, so that we introduce an additional parameter

max senses return only the top k projected senses *after* induction. (Induction operates on all senses.)

Results of the projection to English are summarized in Tab. 3. This was conducted against all dictionaries to predict the English PDTB inventory from all non-DiscMar inventories and multi-source induction compared against the best- and worst-performing direct induction. Considering the average case, multi-source induction without pivot language restriction performs better than average-case single source direct induction for level 0, but not for levels 1-3. However, the pivot

language restriction leads to considerable improvements to the extent that constrained multi-source induction outperforms the best-performing single-source induction. Again, the most effective factor identified was the pivot language restriction, and *only* by means of this restriction, we can outperform single-source direct induction.

As Table 3 indicates, f scores for inferred English discourse relations against the PDTB inventory are comparable with the results achieved before for the English DiscMar inventory for level 0 in the all-dict setting, but substantially higher for level 1. Very likely, this is due to the greater similarity of non-DiscMar inventories with the English PDTB inventory, most of which (except for French) are based on or inspired by PDTB, whereas DiscMar inventories and the French LexConn inventory were independently created and had to be mapped. Clearly, the sense classification of DiscMar is loosely mappable to PDTB, only, also because its categories are underdefined.

4.5. Cross-Lingual Aspects

The best-performing configuration for inducing PDTB senses in the all-dict settings to English is to require the presence of 5 pivot languages prior to induction, with one sense for PDTB level 0, two for level 1, four for levels 2 and 3. To evaluate these parameters and to estimate a cross-linguistically valid threshold value, we now explore induction from *all* languages (except English) into every other language (except English). In scenarios where the maximum number of pivot languages cannot be reached (e.g., for Czech which is only connected via English), we use the maximum number of pivot languages supported by the graph, instead (for Czech, this is thus 1).

Figure 3 shows the number of pivot languages (against best-performing min score thresholds for f, predict one sense only) plotted against f for PDTB level 0. Here, we find that a higher number of pivot languages does *not* necessarily increase prediction quality. The average f (calculated over all languages other than English) is indeed maximal for two pivot languages (closely followed by three and four pivot languages), but drops subsequently. The motivation behind is that a higher number of pivot languages can require longer chains of indirect propagation, and thus to an increase of noise. The medians of best-performing min score thresholds per pivot language restriction are 0.225 (1 pivot language), 0.25 (2 pivot languages), 0.20 (3 to 5 pivot languages), 0.175 (6 pivot languages), and 0.225 (7 pivot languages). As a generalization, we suggest to use 0.20 as (a lower limit for) as min score for indirect induction.

5. Discussion

This paper describes the induction of discourse marker inventories from large-scale lexical networks. Taking a scenario in which a discourse marker inventory is created by translating an existing discourse marker for another language by means of a dictionary as a basis, we found that for languages for which no such bilingual dictionary is available, this can also be achieved by indirect induction over one or multiple pivot languages – if this induction is constrained, e.g., by min pivots threshold that enforces a minimum number of pivot language translations to perform an induction step. In scenarios in which information from discourse marker inventories from multiple source languages can be combined, we find that such a constrained induction can outperform the presumed gold standard of direct induction.

As English seems both the best-connected language in the graph and features multiple discourse marker inventories, a naive approach could be direct induction from English to all other languages. Indeed, this has been tested in Sect. 4.1, but results varied greatly for the dictionaries. As illustrated for the DiscMar inventories (Tab. 1, 2), best results in direct single-source and direct multi-source induction have been achieved with Apertium, and these could be considered gold standard results. By comparison, FreeDict had strong deficits in precision (this is heterogeneous, crowd-sourced material) and MUSE in coverage (as seen from the small number of overall predictions). For languages for which no Apertium dictionary is available, indirect projection in the the all-dicts setting thus represents the only viable choice. For the languages with discourse marker inventories considered here, only Spanish and Catalan have an English Apertium dictionary, but both support PDTB levels 0 and 1 only (from DiscMar).

In indirect induction over all dictionaries, we achieve high levels of recall, although precision is less robust. This indicates that this method is particularly valuable for supporting the semiautomated bootstrapping of discourse marker inventories in which the algorithm produces a number of discourse marker candidates ranked for their respective confidence (predicted level 0 score). Given such data at hand, a language expert can then easily sieve through the top matches of this list. With discourse marker inventories typically containing 50-500 discourse markers only, this allow enable a language expert to create a discourse marker inventory within only a few working hours.

Code and data for our study are published under the Apache v.2 license and available from our GitHub repository.[5] This also includes automatically generated discourse marker inventory stubs for 10 languages (Bulgarian, Greek, Esperanto, Finnish, Japanese, Norwegian, Polish, Russian, Swedish and Turkish) which give discourse marker candidates, their PDTB level 0 score (discourse marker probability), and the associated discourse relations, ranked for their respective score in a simple tabular format with PDTB levels 1-3 merged. Aside from PDTB discourse relations, we generated inventory stubs for RST, using the same al-

[5] http://github.com/acoli-repo/rdf4discourse

dicts	level	min score	min pivot languages	max senses	prediction	p	r	f
best-performing direct induction (over aggregated/all dictionaries, cs,de,fr,it,nl,pt)								
all:pt-en	2				535	0.164	0.815	0.274
all:pt-en	3				707	0.127	0.804	0.220
average scores for direct induction (cs,de,fr,it,nl,pt)								
all	2				604	0.154	0.682	0.242
all	3				645	0.106	0.403	0.164
best-performing pivot language restriction								
all	2	0.50	6	unrestricted	441	0.222	0.632	0.329
all	3	0.75	6	unrestricted	251	0.247	0.369	0.296
best-performing restriction on projected senses								
all	2	0.45	5	4	250	0.364	0.669	0.472
all	3	0.45	5	4	256	0.309	0.622	0.413

Table 3: Inducing the full English PDTB inventory with restrictions on the number of pivot languages, projected senses and min score thresholds

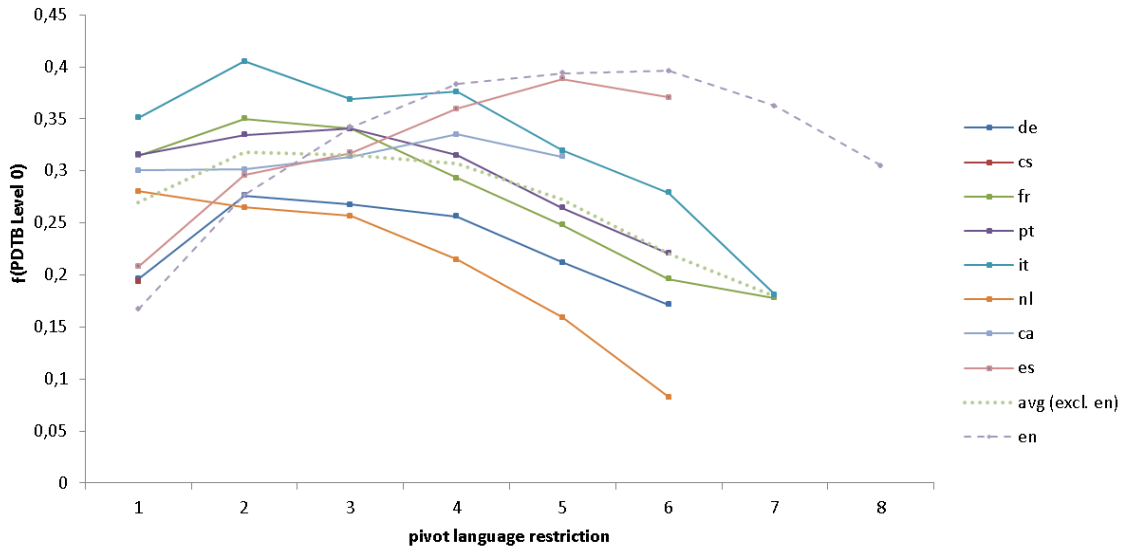

Figure 3: Multi-source/multi-target induction for PDTB level 0 (predict 1 sense only), plotted f over min pivot languages (with best-performing min score from 0.0, 0.05, ..., 1.0).

gorithm and an automated mapping of the original discourse marker inventories retrieved with a SPARQL query from the OLiA Discourse Extensions (Chiarcos, 2014). Furthermore, we also provide inventories with CCR relations (Hoek et al., 2019) constructed in the same way.

In future research, these inventory stubs can be used as input for manual pruning and subsequently be integrated in a community portal for discourse marker inventories such as Connective-Lex (Stede et al., 2019). Prior to doing this, however, we plan to conduct additional experiments with a refined methodology and additional parameters for which the current study provides a baseline. In particular, this includes the addition of corpus information. A natural extension is induction weighted by corpus frequencies (as included in German and one of the three English discourse marker inventories we used, but not in any of the others), but extrapolating corpus frequencies for different discourse relations from corpora annotated for a schema that is not identical (albeit close) to the target classification or estimating them from automated disambiguation is a non-trivial enterprise and beyond the scope of the current experiment.

Finally, we would like to point out a methodological aspect, that is, the application of web standards: Aside from being a contribution to the development of discourse resources and discourse-aware language technology, it is important to note that this was possible only because a large number of resources have been made available beforehand in formats that facilitate subsequent information integration and conjoint querying. The fact that discourse marker inventories, bilingual dictionaries and the underlying inventories of discourse markers are available in RDF, resp., as Linked Data, allows us to easily conduct a large-scale study across multiple families of dictionaries and discourse marker inventories.

6. Acknowledgements

The work described in this paper was conducted at the Applied Computational Linguistics Lab at Goethe University Frankfurt and partially supported by the project "Linked Open Dictionaries (LiODi)", funded as an eHumanities research group by the German Ministry of Education and Research (BMBF, 2015-2022). The induction experiment itself was conducted in the context of the ERC Horizon 2020 Research and Innovation Action "Prêt-à-LLOD. Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors" (2019-2022, grant agreement 825182). The author would like to thank three anonymous reviewers for feedback and input.

7. Bibliographical References

Alonso, L. (2005). *Representing discourse for automatic text summarization via shallow NLP techniques*. Ph.D. thesis, Tesis doctoral. Barcelona: Universitat de Barcelona.

Bański, P. and Wójtowicz, B. (2009). FreeDict: An open source repository of TEI-encoded bilingual dictionaries. In *TEI Members Meeting 2009 (TEIMM-2009)*.

Berners-Lee, T. (2006). Linked data - Design issues. Technical report, W3C.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Bourgonje, P. and Stede, M. (2020). Exploiting a lexical resource for discourse connective disambiguation in German. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5737–5748.

Bourgonje, P., Hoek, J., Evers-Vermeul, J., Redeker, G., Sanders, T., and Stede, M. (2018). Constructing a lexicon of dutch discourse connectives. *Computational Linguistics in the Netherlands Journal*, 8:163–175.

Buch-Kromann, M. and Korzen, I. (2010). The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 127–131.

Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt et al., editors, *Current and New Directions in Discourse and Dialogue*, Text, Speech, and Language Technology; 22, chapter 5. Kluwer, Dordrecht.

Chiarcos, C. and Ionov, M. (2021). Linking discourse marker inventories. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Chiarcos, C., Fäth, C., and Ionov, M. (2020a). The ACoLi dictionary graph. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC-2020)*, pages 3281–3290.

Chiarcos, C., Schenk, N., and Fäth, C. (2020b). Translation inference by concept propagation. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 98–105.

Chiarcos, C. (2014). Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation. In *Proceedings of the 9th Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 4569–4577, Reykjavik, Iceland.

Cimiano, P., McCrae, J., and Buitelaar, P. (2016). Lexicon Model for Ontologies. Technical report, W3C Community Report, 10 May 2016.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data*. Springer.

Da Cunha, I., Torres-Moreno, J.-M., and Sierra, G. (2011). On the development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW-2011)*, pages 1–10, Portland, Oregon, USA.

Danlos, L., Colinet, M., and Steinlin, J. (2015). FDTB1, premiere étape du projet 'French Discourse Treebank': Repŕage des connecteurs de discours en corpus. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, 17.

Das, D., Scheffler, T., Bourgonje, P., and Stede, M. (2018). Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2018)*, pages 360–365, Melbourne, Australia.

Degand, L. (2016). Discourse relational devices in textlink: From (categorical) description to corpus annotation, and back again. In *Discourse Relational Devices (LPTS2016), Univesitat de Valencia*.

Feltracco, A., Jezek, E., Magnini, B., and Stede, M. (2016). LICO: A lexicon of Italian connectives. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 141–145. Accademia University Press.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Fuentes Fort, M. (2008). *A flexible multitask summarizer for documents from different media, domain and language*. Universitat Politècnica de Catalunya.

Gastel, A., Schulze, S., Versley, Y., and Hinrichs, E. (2011). Annotation of explicit and implicit discourse relations in the TüBa-D/Z treebank. In *Multilingual Resources and Multilingual Applications, Proceedings of the Meeting of the German Society of*

Computational Linguistics and Language Technology (GSCL) 2011, pages 99–104.

Hoek, J., Evers-Vermeul, J., Sanders, T. J., et al. (2019). Using the cognitive approach to coherence relations for discourse annotation. *Dialogue & Discourse*, 10(2):1–33.

Huang, H.-H. and Chen, H.-H. (2011). Chinese discourse relation recognition. In *Proc. of 5th International Joint Conference on Natural Language Processing*, pages 1442–1446, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Klyne, G., Carroll, J., and McBride, B. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C Recommendation.

Laali, M. and Kosseim, L. (2014). Inducing discourse connectives from parallel texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING-2014)*, pages 610–619, Dublin, Ireland.

Long, W., Webber, B., and Xiong, D. (2020). TED-CDB: A large-scale Chinese discourse relation dataset on TED talks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2793–2803.

Mann, William C. and Thompson, Sandra A. (1986). Rhetorical Structure Theory: Description and construction of text structures. Technical Report ISI/RS-86-174, Information Sciences Institute.

Mendes, A., del Rio, I., Stede, M., and Dombek, F. (2018). A lexicon of discourse markers for Portuguese – LDM-PT. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC-2018)*, pages 4379–4384, Miyazaki, Japan.

Ordan, N., Gracia, J., Alper, M., and Kernerman, I. (2017). TIAD-2017 shared task – Translation Inference Across Dictionaries. In John P. McCrae, et al., editors, *LDK Workshops 2017: OntoLex, TIAD and Challenges for Wordnets*, Galway, Ireland.

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., and Joshi, A. (2009). The Hindi discourse relation bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 158–161, Prague, Czech Republic.

Pardo, T. A. and Nunes, M. d. G. V. (2004). Relações retóricas e seus marcadores superficiais: Análise de um corpus de textos científicos em português do brasil. *Relatório Técnico NILC*.

Poláková, L., Mírovskỳ, J., Nedoluzhko, A., Jínová, P., Zikánová, Š., and Hajicová, E. (2013). Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (ICJNLP-2013)*, pages 91–99, Nagoya, Japan.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-2008)*, Marrakech, Morocco.

Raymond, C., Riccardi, G., Rodrıguez, K. J., and Wisniewska, J. (2007). The LUNA corpus: an annotation scheme for a multi-domain multi-lingual dialogue corpus. *Decalog 2007*, page 185.

Roze, C., Danlos, L., and Muller, P. (2012). LEXCONN: a French lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, 10.

Scheffler, T. and Stede, M. (2016). Adding semantic relations to a large-coverage connective lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)*, pages 1008–1013.

Stede, M. and Umbach, C. (1998). DiMLex: A lexicon of discourse markers for text generation and understanding. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 1238–1242.

Stede, M., Scheffler, T., and Mendes, A. (2019). Connective-lex: A web-based multilingual lexical resource for connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, 24.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain, July.

Van Der Vliet, N., Berzlánovich, I., Bouma, G., Egg, M., and Redeker, G. (2011). Building a discourse-annotated Dutch text corpus. *Bochumer Linguistische Arbeitsberichte*, 3:157–171.

Versley, Y. (2010). Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–82.

Westphal, P., Stadler, C., and Pool, J. (2015). Countering language attrition with PanLex and the Web of Data. *Semantic Web*, 6(4):347–353.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., and Ogrodniczuk, M. (2019). TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, pages 1–38.

Zikánová, Š., Mírovskỳ, J., and Synková, P. (2019). Explicit and implicit discourse relations in the Prague Discourse Treebank. In *International Con-*

ference on Text, Speech, and Dialogue, pages 236–248. Springer.

A. Lexical Graph Topology

We assume that the topology of dictionary graphs may explain certain differences in the results, therefore, we provide plots of the dictionary collections in isolation and as a group. In these figures, every node represents a language, and nodes with double border represent languages with discourse marker inventories. Every edge indicates the presence of at least one bi-dictionary between these languages. For every node, the shading indicates the number of links to other languages. The exact labels (languages) in these figures may not be readable, but in all cases, a very dark, central node is English.

The Apertium dictionaries (Fig. 4) form a relatively sparse graph which prominently features links between closely related languages (this reflects the application scenario for symbolic translation between related languages). The FreeDict dictionaries (Fig. 5) show a much less regular structure: On the one hand, numerous languages are linked with English only; on the other hand, a small number of widely spoken languages forms a clique of major languages. The topology of the MUSE dictionary graph (Fig. 6) is conceptually similar to that of FreeDict, except that the number of major languages linked with languages other than English is even more restricted. However, while graph topology may have an effect, it cannot be disentangled from other differences between the data sets. As such, Apertium dictionaries are rich linguistic resources created for applications in machine translation, and in this application, they are subject to rigorous evaluation. In particular, these dictionaries have a bias towards providing the most frequent translation only, as rare translations are more likely to be perceived as errors and to be eliminated in the data curation process.

FreeDict dictionaries, on the other hand, are designed for human consumption and created with the goal to *inform* the reader. In that regard, they put a stronger emphasis in providing also less frequent translations, as these may be what a human translator might be looking for when consulting a dictionary.

Finally, MUSE dictionaries are shallow word lists automatically compiled from parallel corpora. As this extraction aimed to extract reliable translation pairs only, this has a similar bias as the Apertium data, but it is not optimized for coverage (like Apertium), but for precision. Moreover, this is not manually curated data, so in general, it will be more noisy.

Figure 7 illustrates the all-dicts graph. However, for readability, we removed all dictionaries (edges) that are not part of a noncyclic path from one discourse marker inventory to another, so these are those relevant for concept propagation between the discourse marker inventories considered here.

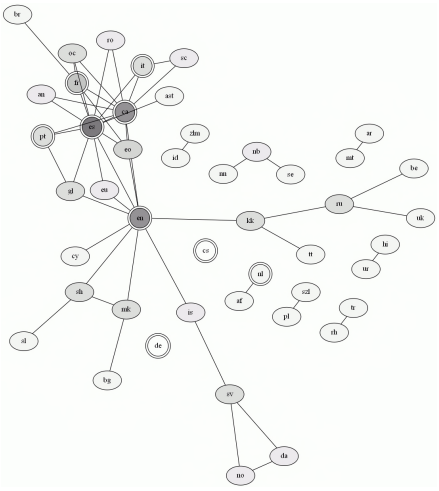

Figure 4: Apertium dictionaries, double lines mark languages with discourse marker inventory

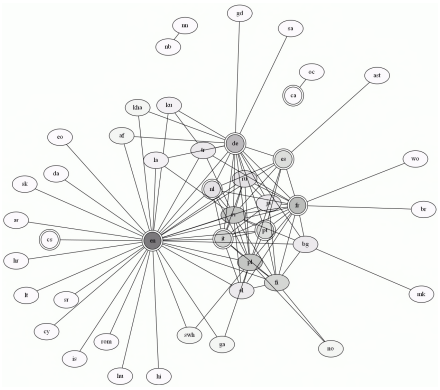

Figure 5: FreeDict dictionaries, double lines mark languages with discourse marker inventory

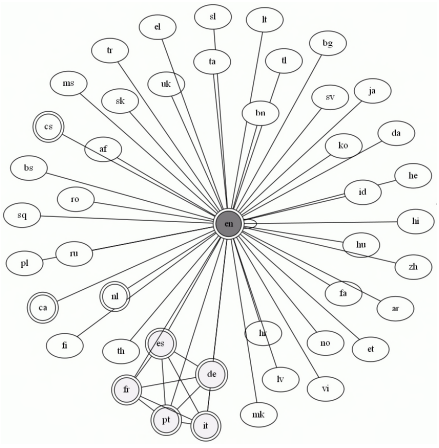

Figure 6: MUSE dictionaries, double lines mark languages with discourse marker inventory

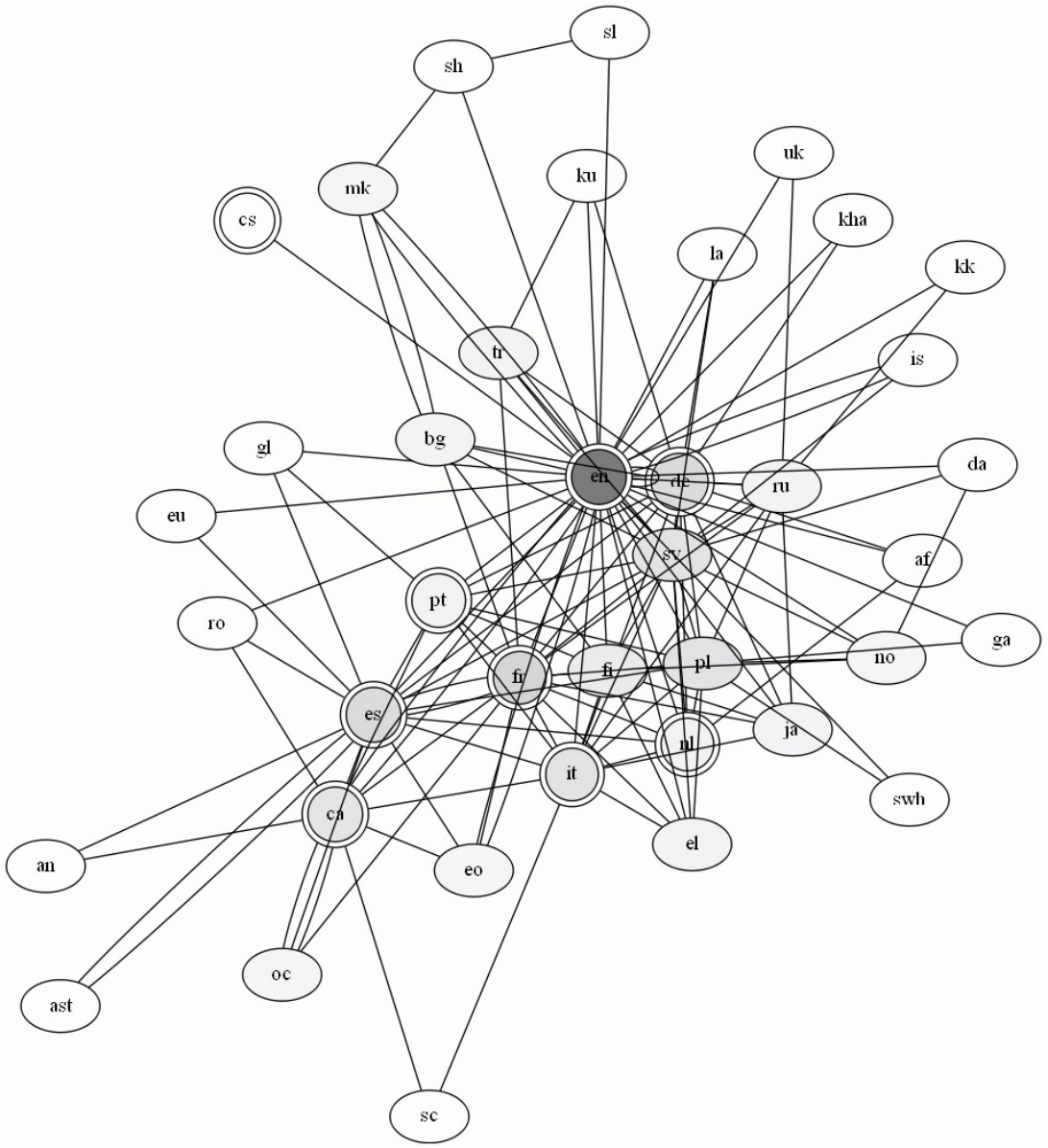

Figure 7: Apertium+FreeDict+MUSE dictionaries conjoint graph, reduced to languages relevant for the evaluation of predicted discourse relations.