View metadata, citation and similar papers at core.ac.uk
brought to you by CO
provided by DSpace at Tartu University L

NEALT PROCEEDINGS SERIES
VOL. 10

Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora

AEPC 2010

December 2, 2010
Tartu, Estonia

Editors

Lars Ahrenberg
Jörg Tiedemann
Martin Volk

NORTHERN EUROPEAN ASSOCIATION FOR LANGUAGE TECHNOLOGY

Proceedings of the workshop on
Annotation and Exploitation of Parallel Corpora

NEALT Proceedings Series, Vol. 10

© 2010 The editors and contributors.

ISSN 1736-6305 (Online)
ISSN 1736-8197 (Print)

Published by
Northern European Association for Language
Technology (NEALT)
http://omilia.uio.no/nealt

Electronically published at
Tartu University Library (Estonia)
http://hdl.handle.net/10062/15893

Volume Editors
Jörg Tiedemann
Lars Ahrenberg
Martin Volk

The publication of the proceedings is supported by the European Regional Development Fund through the Estonian Centre of Excellence in Computer Science (EXCS).

Series Editor-in-Chief
Mare Koit

Series Editorial Board
Lars Ahrenberg
Koenraad De Smedt
Kristiina Jokinen
Joakim Nivre
Patrizia Paggio
Vytautas Rudžionis

Preface

The first workshop on Annotation and Exploitation of Parallel Corpora (AEPC) takes place in Tartu, Estonia on 2nd December 2010 co-located with the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9).

The AEPC workshop brings together researchers that work on parallel corpora for various languages and purposes and features presentations on best practices in annotation and exploration of these corpora for linguistic studies as well as for practical applications.

We received 15 submissions, all of them were reviewed by 3 experts in the field. The reviewing resulted in 9 accepted papers. One accepted paper was dropped because of the author's unavailability for the workshop. The submissions clearly met our expectations for a broad range of topics, and we are happy that many of them report on advances in methodologies, both for manual and automatic corpus annotation. We are also glad that several contributions describe work that ventures into semantic annotation.

We would like to thank all researchers who submitted papers and made this workshop a valuable contribution to our field.

We are also happy to have Matthias Buch-Kromann from the Copenhagen Business School as our invited speaker presenting his work on the Copenhagen Dependency Treebank and new challenges emanating from this research.

We would like to acknowledge the efforts of our program committee helping us to select a good variety of high quality papers.

- Paul Buitelaar (DERI, Galway)
- Anne Göhring (University of Zurich)
- Silvia Hansen (University of Mainz)
- Joakim Nivre (Uppsala University)
- Lonneke van der Plas (University of Geneva)
- Yvonne Samuelsson (Stockholm University)
- John Tinsley (Dublin City University)
- Mats Wirén (Stockholm University)
- Ventsislav Zhechev (Dublin City University)

Furthermore we would like to acknowledge the friendly support by the local organization team at the University of Tartu, in particular Mare Koit, Kaili Müürisep, Kadri Muischnek and Tõnu Tamme for setting up the AEPC web pages, handling the local logistics, and taking care of the proceedings printing. Thanks also to the

TLT chairs Markus Dickinson, Erhard Hinrichs and Marco Passarotti who invited the AEPC workshop as a valuable extension of the TLT workshop series.

We hope that our workshop offers inspiration and ideas for further research and helps to establish new contacts for collaborations in the future.

The publication of these proceedings was supported by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS.

The AEPC organization team

Jörg Tiedemann (Uppsala University)

Lars Ahrenberg (Linköping University)

Martin Volk (University of Zurich)

AEPC Workshop Schedule

Tartu, Estonia

Thursday, 2nd December 2010

9:15-9:30	Welcome: Martin Volk
9:30-10:30	Invited Speaker: Matthias Buch-Kromann: **Open Challenges in Treebanking: Some Thoughts based on the Copenhagen Dependency Treebanks**
10:30-11:00	**Coffee Break**
	SESSION 1 (Chair: Jörg Tiedemann)
11:00-11:25	Xuansong Li, Stephanie Strassel, Stephen Grimes, Safa Ismael, Xiaoyi Ma, Niyu Ge, Ann Bies, Nianwen Xue and Mohamed Maamouri: **Parallel Aligned Treebank Corpora at LDC: Methodology, Annotation and Integration**
11:25-11:50	Maud Ehrmann and Marco Turchi: **Building Multilingual Named Entity Annotated Corpora exploiting Parallel Corpora**
11:50-12:15	Lars Ahrenberg: **Clause restructuring in English-Swedish translation**
12:15-13:45	**Lunch Break**
	SESSION 2 (Chair: Lars Ahrenberg)
13:45-14:10	Natalia Klyueva and David Marecek: **Towards Parallel Czech-Russian Dependency Treebank**
14:10-14:35	Alexandr Rosen: **Mediating between incompatible tagsets**
14:35-15:00	Hans Paulussen and Lieve Macken: **Annotating the Dutch Parallel Corpus**
15:00-15:30	**Coffee Break**
	SESSION 3 (Chair: Martin Volk)
15:30-15:55	Tom Vanallemeersch: **Tree Alignment through Semantic Role Annotation Projection**
15:55-16:20	Yannick Versley: **Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection**
16:20-16:35	Closing Remarks: Jörg Tiedemann

Contents

Open challenges in treebanking: some thoughts based on the Copenhagen Dependency Treebanks

Matthias Buch-Kromann
Copenhagen Business School

E-mail: matthias@buch-kromann.dk

Abstract

Despite the obvious importance and success of treebanks and other linguistically annotated corpora in the current data-driven statistical paradigm in computational linguistics, there are many outstanding challenges. This paper identifies some of the more important challenges, which are mainly concerned with how to exploit synergies by linking up different annotation projects with each other, how to link linguistic annotations to linguistic theory, and how to assess treebank quality without unintended distortions in the way research is conducted in the field.

1 Introduction

Since the creation of the Penn Treebank by Marcus et al [9] in the early 1990's, linguists and computational linguists have succeeded in creating a large number of excellent linguistically annotated corpora (or treebanks, for short). These treebanks cover a large number of languages and a wide range of different linguistic levels, most importantly syntax, part-of-speech, morphology, discourse, coreference, predicate-argument structure, semantics, and word and phrase alignments. They differ from dictionaries and other lexical resources in that they encode linguistic analyses of language phenomena in context, rather than linguistic analyses of words in isolation, and this is the key to their success.

The treebanks created by the field are an important achievement which, together with new statistical techniques, have fuelled the recent paradigm change in computational linguistics from rule-based systems whose language-specific knowledge is encoded as hand-made dictionaries and grammars to data-driven statistical systems whose language-specific knowledge is induced from raw and annotated texts. This paradigm change has spurred the development of a wide range of supervised and semi-supervised statistical techniques in natural language processing that build on annotated corpora — with statistical parsing as the most prominent and successful application so far. Although fully unsupervised techniques have been

proposed in many areas of natural language processing, they have so far mostly failed to produce results that are competitive with supervised or semi-supervised methods, with machine translation as a remarkable exception. Since nothing suggests an imminent breakthrough in fully unsupervised methods outside MT, there is every reason to believe that annotated corpora will continue to play a crucial role as the primary source of language-specific knowledge in most statistical systems in the foreseeable future.

However, despite the obvious importance and success of treebanks in the new statistical paradigm, there are many outstanding challenges. Some of these concern how treebanks are created, some concern how they are utilized, and some how they are compared, merged, and evaluated. In the following, I will describe what I see as some of the main challenges facing the field, and outline some of the tentative steps that have been, or in my opinion should be, taken towards resolving them.

2 The main challenges

Challenge 1: Bridging between different annotation schemes

Current treebanks are based on a rich variety of linguistic theories and annotation schemes. A single, one-size-fits-all annotation scheme covering all languages and linguistic levels is probably neither desirable, nor possible. But the confusion that arises from the current wealth of annotation schemes is a major obstacle in the development of systems that build on treebanks. The problem is particularly acute for systems that need to draw on several different treebanks simultaneously. For example, Chiang [6] identifies badly interacting source and target language annotation schemes as an important obstacle in supervised tree-to-tree translation, and Meyers [13, 11] points to the lack of coordination between annotation projects as a major obstacle in annotation merging.

An example of the kind of confusion that may arise, even between closely related schemes, is illustrated by Figures 1 and 2, which show the syntactic analysis of four different constructions in two dependency-based annotation schemes: the dependency conversion of the Penn Treebank produced by the PennConverter [8], and the native dependency annotation used in the 100,000 word English part of the Copenhagen Dependency Treebanks [5, 3], the second-largest native (non-converted) dependency treebank for English after the Prague English Dependency Treebank [7]. Although the two annotation schemes are both dependency-based and therefore fall within the same broader family of annotation schemes, the specific differences between the analyses are considerable (non-shared unlabeled arcs are shown as dotted red arcs). The PennConverter scheme takes a semantics-oriented view, motivated by its intended use as training material for parsing semantic dependencies, where content words (nouns, main verbs) tend to dominate function words (determiners, auxiliary verbs, prepositions). In contrast, the CDT scheme takes a syntax-oriented view, where function words tend to dominate content words. There are other differences as well, eg, with respect to the analysis of

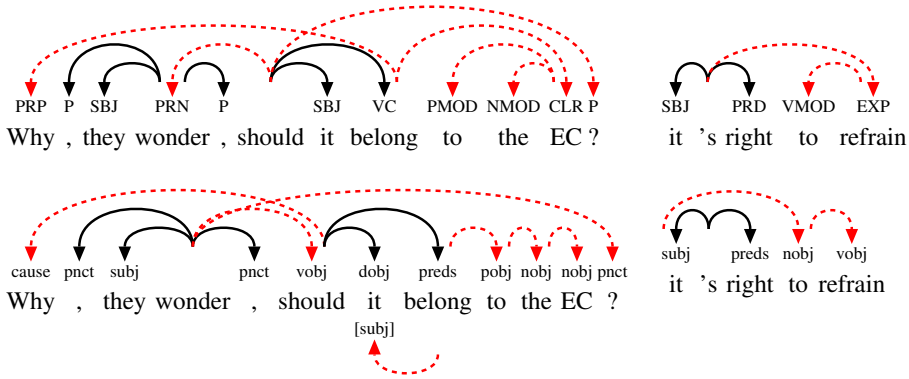

Figure 1: Differences between PennConverter's dependency conversion of the Penn Treebank (top) and the CDT scheme (bottom): attribution (left) and expletive construction (right). (Differing arcs shown with dashed lines.)

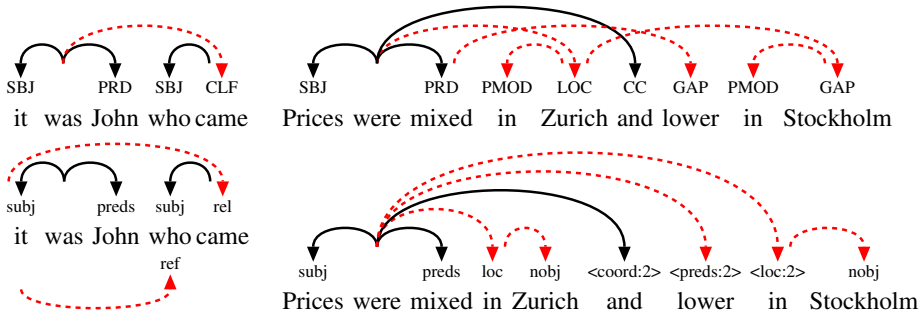

Figure 2: Differences between the PennConverter scheme (top) and the CDT scheme (bottom): cleft sentence (left) and gapping (right)

attribution verbs, gapping, extrapositions, cleft sentences, and the use of additional secondary dependencies and coreference links in the CDT treebanks.

Both schemes are linguistically well-motivated, so it is not a question about one scheme being right and the other being wrong. Indeed, with differences like these, there is no objective criterion for deciding which annotation framework provides the most empirically adequate analysis of the texts. From a theoretical point of view, it may even be misleading to talk about the best scheme, as if there is only one: since we are modelling unobservable properties of language, there is no guarantee that we cannot end up with a set of highly different models which are equally adequate with respect to their observable consequences. For this reason, the relevant challenge is not to create a single unified annotation scheme to be used by all treebanks created in the field, an impractical and unrealistic task — the relevant challenge is to find better ways of translating between different annotation schemes and merging them, in order to pool costly treebank resources.

There are probably many ways of achieving this goal, but the diagram in Fig-

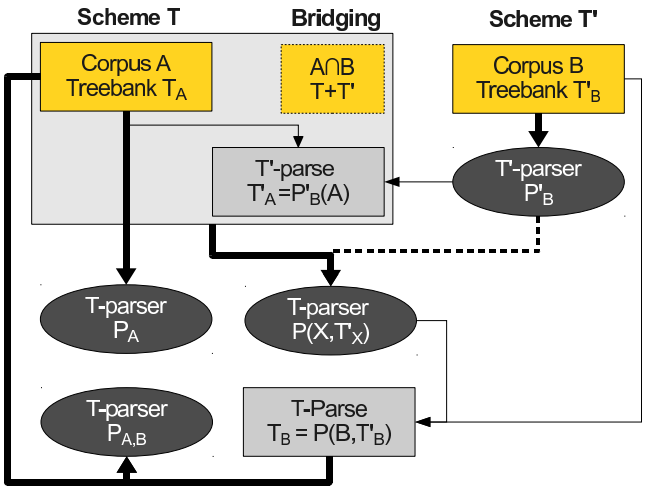

Figure 3: Treebank conversion: Converting a T'-treebank T'_B to a T-treebank T_B, and creating T-parsers $P(X,T'_X)$ and $P_{A,B}$ by pooling treebanks.

ure 4 illustrates the kind of system that I have in mind for treebank conversion and automatic annotation (for simplicity, I will refer to a linguistically annotated corpus T as a 'treebank', and to an automatic annotation system P trained on T as a 'parser'). In the diagram, I am assuming that we have two annotation schemes T,T' applied to two corpora A,B respectively, yielding manually corrected "gold" treebanks T_A,T'_B. For example, T_A might be the Copenhagen Dependency Treebank for English, T'_B might be the Penn Treebank, and my goal might be to convert the Penn Treebank to CDT format, or to build a better CDT parser that would use the Penn Treebank as training material.

We can create parsers P_A,P'_B without any pooling by training them on T_A,T'_B independently. But in order to pool the treebanks, we need to do something else. Inspired by stacked dependency parsing,[1] one possibility is to use the P'_B parser to create an automatically parsed T'-corpus T'_A, and use T_A,T'_A to create a stacked T-parser $P(X,T'_X)$ where T'_X is a T'-annotation of X, eg, a parse produced by P'_B. This stacked parser can then be used to convert T'_B into a T-treebank $T_B = P(B,T'_B)$. Finally, the original treebank T_A and the converted treebank T_B can be pooled to train a new T-parser $P_{A,B}$. The parser $P_{A,B}$ and the stacked parser $P(X,P'_B(X))$ will utilize the information from both T_A and T'_B, and can therefore be expected to perform better than the P_A parser, especially if T_A is a large high-quality treebank. When using the stacked parser $P(X,T'_X)$ as a conversion system, it would probably be helpful to include a designated *bridging treebank* in the system, ie, an overlapping subcorpus $A \cap B$ where the T' annotation has been hand-converted into the

[1] Stacked dependency parsing [16, 10, 18], which is the current state-of-the-art approach in dependency parsing, employs two or more dependency parsing systems to a single treebank, so our setup is a slight variation of the original setup.

corresponding T annotation.[2]

As a variation on this setup, the corpus B might consist of parallel texts, with A as a subcorpus of B, and T'_B, T_A as treebank annotations of the source and target texts in B and A, respectively; $T_A + T'_A$ will then be available as a bridging treebank. The goal is to extend the T-annotation to B. For example, in the CDT treebanks, T'_B might be the 100,000 word treebank for the source language Danish, and T_A might be one of the 30,000 word treebanks for the target languages English, German, Italian, or Spanish, and the goal is to extend the target language treebank to all translations of texts in the Danish treebank. In this case, we would like to train a stacked parser $P(X, T'_X)$ from $T'(A)$ and the bridging treebank $T + T'$ on A, and use the parser to produce $T_B = P(B, T'_B)$.

To sum up, the challenge is to make it easier to convert annotations from one scheme to another, and to create automatic annotation systems that can utilize multiple treebanks with different annotation schemes simultaneously. Solving this challenge will have great practical value, because it will make it easier to convert one treebank format into another, transfer treebank annotations from one language to another, and create monolingual and synchronous parsing systems that build on pooled treebank resources. Small, high-quality bridging treebanks can be expected to improve the quality of these systems, and building bridging treebanks and determining how large they need to be, is therefore an important task for future research.

Challenge 2: Bridging annotations at different linguistic levels

Many annotation projects take a narrow scope where they focus on a single language, a single linguistic level, and perhaps even a single text genre. Many of the most influential annotated corpora for English are based on a narrow-scope approach: eg, Penn Treebank, Penn Discourse Treebank, PropBank, NomBank, TimeBank, and the MATE/GNOME scheme for coreference annotation. Other annotation projects take a wide scope where they seek to provide a coherent unified annotation for a wide variety of languages, linguistic levels, and text genres: eg, the Prague Dependency Treebanks (Czech, English, Arabic), the Copenhagen Dependency Treebanks (Danish, English, German, Italian, Spanish), and OntoNotes (English), which cover syntax, morphology, semantics, discourse, and coreference.

From a scientific perspective, the narrow-scope and wide-scope approaches complement each other: narrow-scope encourages deep explorative analysis of a narrowly defined set of phenomena, whereas wide-scope encourages a focus on the integration between the different linguistic levels, including their interfaces, similarities between the different levels, and their link to a unified linguistic theory. The lack of coordination between narrow-scope treebanks means that they may be based on mutually incompatible assumptions about the underlying linguistic structure, the division of labour between the different treebanks, and the choice of analyses for the phenomena where they overlap. This can make it difficult to

[2]The English CDT treebank includes a small 4,500 word CDT-annotated subset of the Penn Treebank, which can be used as a bridging treebank by treebank conversion systems.

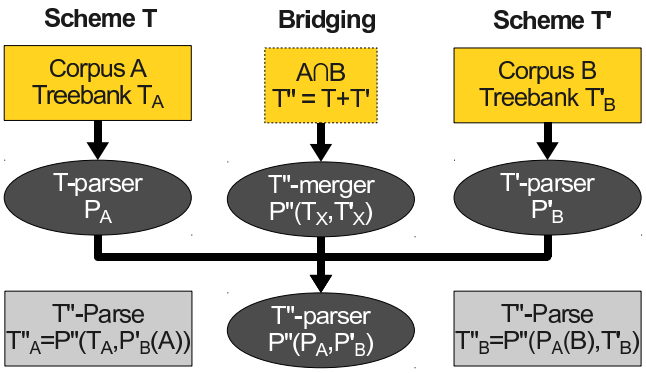

Figure 4: Annotation merging: Merging treebanks T_A, T'_B using parsers P_A, P'_B and merger P'' trained on gold-standard merge $T''_{A\cap B}$, resulting in T''-parser P''.

produce a coherent unified wide-scope treebank by merging several narrow-scope treebanks for different linguistic levels, as pointed out by Meyers et al [13, 11].

Since future applications in language technology are likely to require a coherent set of annotations at several linguistic levels, the integration between annotations at the different linguistic levels should be a key priority in future treebanking research — either as research in unified wide-scope annotation, or as research in systems for annotation merging. The GLARF system proposed by Meyers et al [11] is partly rule-based and designed for a specific set of treebanks. It would therefore be desirable with research in probabilistic general-purpose annotation merging systems that could be trained on two (or more) treebanks T_A, T_B on the basis of a small bridging treebank $T_{A\bigcap B}$, as shown schematically in Figure 4.

Challenge 3: Building a multi-parallel "bridging" corpus community

The researchers in the field should agree on a balanced, general-purpose, mixed-genre English corpus to be used as the English component in a collaborative multi-parallel "bridging" corpus. National research groups would be responsible for translating the English corpus into their own language and contributing their annotations. If these translations formed a substantial part of the corpora used in national treebank projects, it would be a lot easier to merge annotation systems and transfer annotations from one language to another.[3]

The English source corpus should be constructed with great care. It should be composed so that it is suitable for annotation at all linguistic levels (morphology, syntax, discourse, anaphora, semantics), with a permissive license (preferably an open-source license) that places as few restrictions on the subsequent use as pos-

[3]Translations are known to be coloured by the source language, so national treebank projects cannot be expected to work on translation corpora exclusively. But since translation to and from English is one of the major applications for language technology, the decision to include translations as a substantial part of the annotated corpus would make sense, also from a purely national perspective.

sible. To accommodate the needs of different treebank projects large and small, the English source corpus should be diverse, balanced, mixed-genre and general-purpose, and structured as an onion: there should be a tiny core corpus (say, 1,000 words) consisting of isolated sentences or small text excerpts chosen for their linguistic variation, supplemented with a wide range of larger corpora (say with 3k, 10k, 30k, 100k, etc. up to 10M words) that extend the smaller corpora by expanding the existing text excerpts and adding new text excerpts from more texts.[4]

It is important that the bridging corpus is accepted by the researchers in the field as the standard base corpus in most annotation projects (unless there is good reason otherwise), so the decision procedure must be thought out carefully. Eg, perhaps a community committee can specify a set of desiderata for the English source corpora. Different research groups can then come up with competing proposals, which the entire treebanking community can vote on — with a possible revote after a phase where the best proposals are merged.

Challenge 4: Linking treebanks with linguistic theories

When creating a treebank, the linguistic annotations are not observable quantities, but theoretical constructs informed by the annotators' conception of linguistic theory coupled with their intuitions about the language and the text. To be meaningful, annotations must therefore be interpreted within some notion of linguistic theory. We obviously need to be careful: we cannot corroborate a theory with annotations that presume the theory, which is why some annotation projects seek to be "theory-neutral", ie, to avoid basing the annotations too closely on linguistic theory. On the other hand, a theory-neutral approach is not bullet-proof, and it takes more than a few counter-examples in a theory-neutral annotation to disprove a theory: after all, the counter-examples might just be artifacts of ill-conceived annotation guidelines or misjudged intuitions by the annotators.

From a methodological point of view, I think the best solution to this dilemma is to give the annotations a clear interpretation in terms of linguistic theory, but allow annotators to mark cases where the theory is hard to apply. During the annotation, this will promote a rich interaction between the annotation and the linguistic theory. Moreover, without a clearly formulated interpretation of the annotations, it is difficult for other researchers to criticize the annotations or the underlying theory. The Copenhagen Dependency Treebanks may serve as an example of the approach that I am advocating. The annotations are heavily informed by the dependency theory Discontinuous Grammar [2], which stipulates how dependency structures determine the word order and the compositional semantics. The linguistic theory has been a huge help in the design of the annotation scheme. Since the theory is much cleaner if discourse structure is viewed as the continuation of sentential syntax to the inter-sentential level, a tree-structured discourse supplemented with coreference relations has been a guiding principle in our discourse annotation (a

[4] The "Pie in the Sky" corpus [12] may serve as an inspiration.

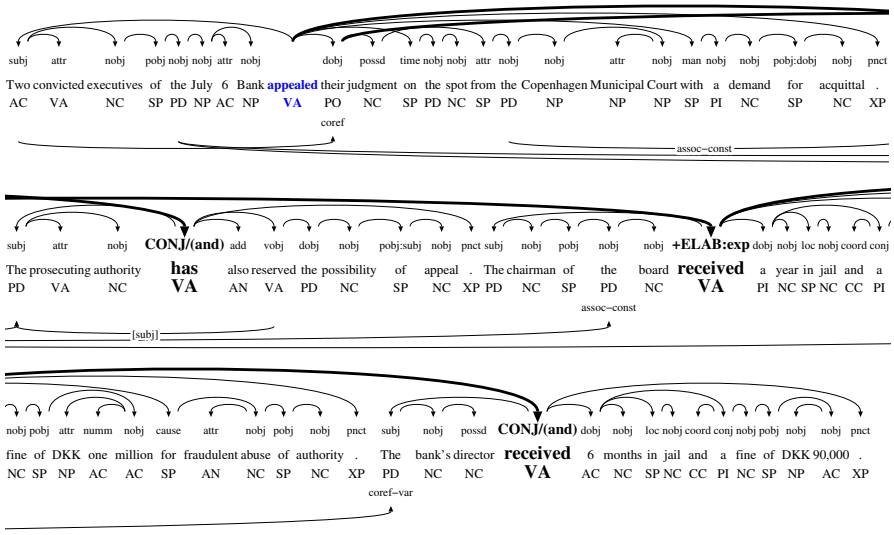

Figure 5: Example of a CDT syntax-discourse annotation. The primary dependencies form a tree structure shown on the top, the bottom arcs encode additional secondary dependencies and coreference links needed for semantic interpretation.

similar principle applies to morphology, with some adjustments that deal with non-concatenative morphology [5, 14]). The annotators have encountered hard cases, which have sharpened the theory by requiring revisions in theory and analysis. But the basic assumption about a tree-based discourse structure has held up, which has allowed us to contribute to the long-standing theoretical debate about whether discourse structure is best viewed as a tree or as a graph [4]. Figure 5 shows a unified CDT annotation of syntax and discourse.

The new data-driven statistical paradigm in computational linguistics has many virtues, but it has also led to an unhealthy decline in interactions with theoretical linguistics: treebanks and other linguistically annotated corpora are now the only place where linguistics really comes into play. In my view, one of the main challenges for the treebanking community is to build a stronger two-way interaction between linguistic annotation and linguistic theory: the theory should provide sense and direction for the annotations as well as their application in natural language processing. This is the best way of using linguistic theory as a guide to simpler and more useful annotation, the best way of moving linguistic theory forward, and the best way of bringing linguistics back into computational linguistics.

Challenge 5: Quantifying treebank quality

Any treebank project is faced with a wealth of choices where there are good linguistic arguments for more than one choice. Even close collaborators working within the same annotation project may disagree about the right analysis. The wealth of

annotation schemes in treebanking is therefore unsurprising. Ideally, these choices should be made in a more principled way, and in some distant future, treebank quality will perhaps be evaluated by the perplexity that a linguistically based language model trained on the treebank assigns to an unseen corpus, and hard design decisions in the annotation scheme will perhaps be made partly or fully on the basis of such scores. But so far, there is no easy solution in sight. Measuring treebank quality is probably one of the hardest and most important outstanding problems in the field, and any research that can adress these problems even tentatively should be encouraged by the field.[5]

Unfortunately, in the absence of a better measure, inter-annotator agreement seems to have taken up an unhealthy role as the primary measure of annotation quality. Reviewers routinely request agreement figures when reviewing treebank papers, and people have suggested that low κ values (say, below 0.8 or 0.67) makes inter-coder reliability so low as to leaving the annotations useless (cf. the excellent review article by Artstein and Poesio [1]). There is no doubt that agreement and related measures have important uses: annotation projects should keep constant track of inter-annotator agreement and confusion tables for the individual relations, and prompt annotators and linguists to try to eliminate major sources of disagreement. However, apart from that, it is not clear that agreement has a constructive role to play: Reidsma and Carletta [17] have shown that agreement is a poor predictor of the performance of a system based on machine learning; more importantly, if used as a proxy for annotation quality by treebank designers and reviewers, an exaggerated focus on agreement may lead to distortions in the way treebanks are designed.

Agreement and confusion scores are highly beneficial when used to identify misunderstandings and formulate linguistically well-motivated clarifications of the annotation scheme. The distortions happen when treebank projects design their annotation schemes so that they optimize agreement, regardless of whether the agreement-boosting measures fail to be linguistically motivated. As an extreme case, we can construct a dependency treebank with 100% inter-annotator agreement by picking an annotation scheme where every word is analyzed as a dependent of the preceding word, using a single dependency label; a parser trained on this treebank would have a 100% labeled attachment score. Slightly more subtle agreement measures that correct for chance agreement, like κ and α, will assign a low agreement score to this annotation scheme; but even they can be tricked to yield a near-100% score, if we subdivide our single label into several subtypes which are easy for humans (and parsers) to disambiguate. For example, we can use the word class (or some other easily inferable quantity) as our relation label — which is what the Copenhagen Dependency Treebank, and many other treebanks, inadvertently do when they use labels such as 'pobj' for prepositional objects, 'nobj' for nominal objects, etc. For this reason, chance-corrected agreement provides a false sense of

[5] Nilsson, Nivre and Hall [15] have made an interesting experiment where they show that simple treebank transformations (changing the structure of coordinations or verb groups) can improve MaltParser performance, but they also show that these transformations are suboptimal for the MSTParser, ie, their method cannot really be used to make an unambiguous case for one analysis over the other.

security, and does not produce comparable scores even when comparing treebanks with the same underlying corpus and number of labels.

Another way of boosting agreement is to instruct annotators to always use a particular ad-hoc analysis for ambiguous constructions, eg, preferring VP attachments over NP attachments when in doubt. This move increases agreement, but also induces an arbitrary bias which may be harmful in some applications. Semi-automatic annotation also boosts agreement because annotators are biased towards accepting the analysis proposed by the parser, unless it is clearly wrong. Even worse, we may decide to simplify the annotation scheme by merging linguistically well-motivated labels that are often confused with each other, such as dependency relations that reflect fine-grained semantic distinctions (adverbial relations, discourse relations), and so on. The experience of the CDT annotators, and many others in the field, is that semantic distinctions are really hard to make, and that disagreements are often caused by truly ambiguous texts where the two differing analyses either lead to essentially the same meaning, or the context does not contain sufficient information about the speaker's true intentions. But that does not necessarily imply that the distinction does not encode important information, it is just noisy information.

The big question is: do we really improve treebank quality by making linguistically unmotivated design decisions that improve agreement, but reduce informativity? There are many desiderata for a good measure of treebank quality, but one of the more important is that it should be impossible to improve the quality score by merely throwing away information, for example, by merging labels mechanically. Agreement clearly fails on this criterion. Perhaps chance-corrected agreement measures can be fixed in part by measuring agreement using the highest-scoring set of merged labels, rather than the unmerged set of labels, but it is not clear that the resulting score would be interesting.

Until we get a better measure of treebank quality, reviewers should probably focus less on total agreement and more on a qualitative assessment of the confusion table, which encodes the probability $\mathrm{Conf}(l'|l)$ with which the other annotators used label l' when one of the annotators used label l. In the CDT project, 10% of the annotated texts are double-annotated: this allows us to compute a confusion table, which is included in the CDT annotation manual [3]. As an illustration, Figure 6 shows some of the confusion scores for the syntactic relations in the ongoing CDT annotation: "Agr" specifies the relation-specific agreement, ie, how often did the annotators agree on the label when one of them used the label; "N" specifies the number of tokens for which one of the annotators used the label; and "SN1" specifies the primary signal-to-noise ratio, defined as the ratio between the probability that the other annotators used the same label relative to the probability that the other annotators used the most frequent alternative label. In a classification task, the SN1 ratio can be expected to show a better correlation with machine learning success than agreement, since most classifiers will pit the two highest-ranked labels against each other, ie, the label will be hard to learn if the ratio is smaller than 1. Perhaps an even better predictor of classification success can be constructed by

Rel	Agr	SN1	N	Confusion list
expl	86%	8.6	53	expl$_{86\%}$ subj$_{10\%}$ preds$_{1\%}$ time$_{0\%}$ pobj$_{0\%}$
focal	42%	3.2	38	focal$_{42\%}$ attr$_{13\%}$ other$_{8\%}$ pnct$_{6\%}$ loc$_{5\%}$ nobj$_{3\%}$ err$_{2\%}$ correl$_{2\%}$ eval$_{2\%}$ mod$_{2\%}$ pobj$_{2\%}$ subj$_{1\%}$ dobj$_{1\%}$ ...
iobj	63%	2.4	19	iobj$_{63\%}$ dobj$_{26\%}$ robj$_{5\%}$ pnct$_{1\%}$ subj$_{1\%}$ nobj$_{0\%}$ attr$_{0\%}$ possd$_{0\%}$ modp$_{0\%}$
conc	21%	1.8	23	conc$_{21\%}$ contr$_{13\%}$ mod$_{13\%}$ prg$_{8\%}$ other$_{8\%}$ pobj$_{6\%}$ nobj$_{6\%}$ conj$_{5\%}$ attr$_{5\%}$ pnct$_{4\%}$ dobj$_{2\%}$ subj$_{2\%}$ possd$_{1\%}$ appr$_{0\%}$
iter	19%	0.4	26	time$_{46\%}$ iter$_{19\%}$ other$_{7\%}$ vobj$_{5\%}$ attr$_{3\%}$ eval$_{3\%}$ mod$_{3\%}$ nobj$_{1\%}$ dobj$_{1\%}$ relr$_{1\%}$ cause$_{1\%}$ name$_{1\%}$ ...

Figure 6: Some confusion scores from the CDT annotation manual.

taking the frequencies with which the different labels occur into account.

Although the CDT annotation is still ongoing and we hope to improve the inter-annotator consistency, it is worth noting that even labels with a high degree of confusion contain a lot of information, which will be lost if we start merging labels to improve agreement. By releasing the confusion table along with the treebank, the decision about which labels to merge can be left to the users of the treebank. However, when allowing a higher level of disagreement in the treebank, we also have to reconsider how we score parsers trained on the treebank. That is, the parser must get a score of 1 if it produces time or iter when the gold standard says iter, but a score of 0 if it produces subj. For example, a parser that produces label l' when the gold standard has label l can reasonably be scored with:

$$\max\left(1, \frac{\text{Conf}(l'|l)}{\text{Conf}(l|l)}\right)$$

It is quite possible that there are better ways of using the confusion table to score data-driven systems. The central point here is that the current focus on agreement in treebanking is unfortunate because it has unintended side effects in terms of what people decide to annotate and how they design their annotation schemes: increasing agreement by increasing bias with ad-hoc rules or losing informativity is not necessarily what we need most at the present state of our science.

3 Conclusions

With the advent of data-driven systems, linguistic annotation has become a great success and is maturing as a field. There is however still a number of important unsolved challenges. Most of them are concerned with how to exploit synergies by linking up different annotation projects with each other, even when they use different base corpora, focus on different linguistic levels and different languages, and are based on different conceptions of linguistic theory. Designated data-driven

bridging tools coupled with collaborative bridging corpora are probably the key to long-term success in this area.

The wealth of different annotation schemes suggest that we need more research in how we assess the quality of linguistic annotations, and how we compare competing annotation schemes. At the same time, we must be careful to avoid that our measures of annotation quality do not lead to unintended incentives that distort what people annotate and how they design their annotation schemes. In particular, the current focus on inter-annotator agreement is probably unfortunate because the scores are hard to compare and encourage information loss and bias in the form of linguistically unmotivated ad-hoc principles. Finding better measures of annotation quality is therefore a key priority for the field.

4 Acknowledgments

My research was supported by a grant from the Danish Research Council for the Humanities. Thanks to my colleagues from the Copenhagen Dependency Treebank Project for many inspiring discussions, and to the workshop organizers for their valuable help.

References

[1] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(3), 2008.

[2] Matthias Buch-Kromann. *Discontinuous Grammar. A dependency-based model of human parsing and language learning*. VDM Verlag, 2009.

[3] Matthias Buch-Kromann, Morten Gylling-Jørgensen, Lotte Jelsbech Knudsen, Iørn Korzen, and Henrik Høeg Müller. The Copenhagen Dependency Treebank repository. http://code.google.com/p/copenhagen-dependency-treebank, 2010.

[4] Matthias Buch-Kromann, Iørn Korzen, and Daniel Hardt. Syntax-centered and semantics-centered views of discourse. Can they be reconciled? In *Proceedings of the DGfS 2011 workshop Beyond Semantics (to appear)*.

[5] Matthias Buch-Kromann, Iørn Korzen, and Henrik Høeg Müller. Uncovering the 'lost' structure of translations with parallel treebanks. In Fabio Alves, Susanne Göpferich, and Inger Mees, editors, *Methodology, Technology and Innovation in Translation Process Research*, volume 38 of *Special issue of Copenhagen Studies of Language*, pages 199–224. 2009.

[6] David Chiang. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1443–1452, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[7] Silvie Cinková, Josef Toman, Jan Hajič, Kristýna Čermáková, Václav Klimeš, Lucie Mladová, Jana Šindlerová, Kristýna Tomšů, and Zdeněk Žabokrtský. Tectogrammatical annotation of the Wall Street Journal. *Prague Bulletin of Mathematical Linguistics*, 92, 2009.

[8] Richard Johansson and Pierre Nugues. Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA 2007*, 2007.

[9] M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The Penn Treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*, 1994.

[10] André F. T. Martins, Dipanjan Das, Noah A. Smith, and Eric P. Xing. Stacking dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 157–166, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[11] A. Meyers, M. Kosaka, N. Xue, H. Ji, A. Sun, S. Liao, and W. Xu. Automatic recognition of logical relations for English, Chinese and Japanese in the GLARF framework. In *SEW-2009 at NAACL-HLT-2009*, 2009.

[12] Adam Meyers. Introduction to Frontiers in Corpus Annotaton II: Pie in the Sky. In *Proc. of ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*. 2006.

[13] Adam Meyers. Compatibility between corpus annotation efforts and its effect on computational linguistics. In Paul Baker, editor, *Contemporary Approaches to Corpus Linguistics*. Continuum Publishers, 2009.

[14] Henrik Høeg Müller. Annotation of morphology and NP structure in the Copenhagen Dependency Treebanks. In *The Ninth International Workshop on Treebanks and Linguistic Theories*, 2009.

[15] Jens Nilsson, Joakim Nivre, and Johan Hall. Generalizing tree transformations for inductive dependency parsing. In *Proc. ACL-2007*, 2007.

[16] Joakim Nivre and Ryan McDonald. Integrating graph-based and transition-based dependency parsers. In *Proceedings of ACL-08: HLT*, pages 950–958, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[17] D. Reidsma and J. Carletta. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326, 2008.

[18] Anders Søgaard and Christian Rishøj. Semi-supervised dependency parsing using generalized tri-training. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1065–1073, Beijing, China, August 2010.

Parallel Aligned Treebank Corpora at LDC: Methodology, Annotation and Integration

Xuansong Li, Stephanie Strassel, Stephen Grimes, Safa Ismael, Xiaoyi Ma, Niyu Ge, Ann Bies, Nianwen Xue, Mohamed Maamouri

Linguistic Data Consortium, IBM, Brandeis University
Email:{xuansong,strassel,sgrimes,safa,xma,bies,maamouri}@ldc.upenn.edu, niyuge@us.ibm.com, xuen@brandeis.edu

Abstract

The interest in syntactically-annotated data for improving machine translation quality has spurred the growing demand for parallel aligned treebank data. To meet this demand, the Linguistic Data Consortium (LDC) has created large volume, multi-lingual and multi-level aligned treebank corpora by aligning and integrating existing treebank annotation resources. Such corpora are more useful when the alignment is further enriched with contextual and linguistic information. This paper details how we create these enriched parallel aligned corpora, addressing approaches, methodologies, theories, technologies, complications, and cross-lingual features.

1 Introduction

Parallel aligned treebank (PAT) refers to sentence-aligned data annotated with morphological/syntactic structures and aligned manually or automatically at one or more sub-sentence levels, such as the Japanese-English-Chinese PAT (Uchimoto et al. [7]) or the English-German-Swedish PAT (Volk et al., [8]). Incorporating contextual/linguistic information into a PAT is a new trend, opening up new possibilities for reducing word alignment error rate (Ittycheriah et al. [2]) and enhancing translation quality in statistical machine translation (SMT) models. One such effort is the incorporation of contextual features into tree-alignment (Tiedemann et al. [6]). As a part of this trend, LDC is now manually aligning Penn treebanks. To enrich the word-level alignment, a layer of tagging annotation is incorporated into the alignment to capture contextual and cross-lingual features. Focusing on Arabic, Chinese, and English, LDC has produced a large amount of PAT data as shown in Figure 1.

	Arabic-English PAT				**Chinese-English PAT**				
Genre	***Arb-w***	***Token***	***En-w***	***Seg***	***Ch-w***	***Char***	***En-w***	***Ctb-w***	***Seg***
NW	198558	290064	261303	8322	160477	240920	164161	145925	5322
BN	201421	259047	266601	12109	---	---	---	---	---
BC	---	---	---	---	117630	176448	91650	122714	7156
WB	19296	28138	26382	853	86263	129594	89866	82585	3920
Total	419275	577249	554286	21284	364370	546962	345677	351221	16398

Figure 1: Data Profile

In the above chart, NW, BN, BC, and WB stand for newswire, broadcast news, broadcast conversation, and web data, "Arb-w" for Arabic source words, "En-w" for English words, "Ch-w" for Chinese words, "Char" for Chinese characters, "Ctb-w" for Chinese treebank words, "Token" for tokenized tokens, and "Seg" for segmented sentences. Chinese words are based on characters/1.5.

The most common practice of creating a PAT corpus is to align existing treebank data. Such treebank resources provide mono-lingual syntactic annotations on tokens produced by a particular tokenization scheme. The alignment annotation begins with these leaf tokens to produce ground/base level alignment upon which higher-level alignment can be automatically induced. The optimal ground/base level alignment should be based on the minimum translation unit. In the context of parallel alignment, the minimum translation units refer to context-free atomic semantic units during translation. In this paper, we call it a *linear* approach if the tree leaf tokens are used as the minimum translation unit for alignment. Unfortunately, the tokens used for treebank annotation may not always be the desired minimum tokens for ground/base level alignment. Then the *non-linear* approach would call for another tokenization scheme (other than the treebank tokenization) to produce minimum translation tokens. At LDC, we create the Arabic-English PAT following the *linear* approach, and the Chinese-English PAT following the *non-linear* approach.

The paper is laid out as follows: Sections 2 and 3 discuss data source and tokenization issues respectively; Section 4 elaborates on alignment and tagging annotation at LDC; Section 5 introduces treebanks used for LDC PAT corpora; Section 6 presents the data structure of a PAT; Section 7 describes complications and challenges in creating a PAT; Section 8 concludes the paper.

2 Data Source

Source data used for PAT corpora are harvested by LDC in four genres: newswire, broadcast news, broadcast conversation, and web. Source Arabic and Chinese data are collected from various TV/broadcast programs (Figure 2). Web data are newsgroups and weblogs from on-line resources. The harvested data are manually segmented into sentences by LDC, which are further outsourced to professional translation agencies to produce high quality English translation data.

Language	***Source of Programs***
Arabic	Agence France Presse, Al-Ahram, Al Hayat, Al Quds-Al Arabi, An Nahar, Asharq Al-Awsat, Assabah, Al Alam News Channel, Al Arabiyah, Al Fayha, Al Hiwar, Al Iraqiyah, Al Ordiniyah, Bahrain TV, Dubai TV, Oman TV, PAC Ltd., Saudi TV, Syria TV, Aljazeera.
Chinese	China Military Online, Chinanews.com, Guangming Daily, People's Daily Online, Xinhua News, China Central TV, 2005 Phoenix TV, Sinorama magazines.

Figure 2: Data Sources

3 Tokenization and Segmentation

Raw data need to be tokenized and/or segmented for alignment and treebank annotation. When a PAT corpus is created with the *non-linear* approach, another tokenization scheme needs to be defined for the base-level alignment. With the *linear* approach, no further tokenization scheme is needed. Both of the approaches directly extract leaf tokens from existing parallel treebank data. The extracted tokens may or may not be the smallest translation units for alignment. For our PAT, we use the extracted English and Arabic tokens as the minimum translation units for base-level alignment while the extracted Chinese tokens cannot serve as base-level alignment tokens because some of them need to be further split in order to become minimum translation units.

The English tokens are leaves from the Penn English Treebank. The tokenization has the following features: words separated by white spaces, contractions split, punctuations separated from surrounding words, and the apostrophe (‘,‘s) treated as a separate token. Most hyphens are separate tokens while some are treated as part of words.

Arabic tokenization/segmentation is complex due to the rich morphological features of Arabic. Arabic treebank tokenization splits clitics (except “determiner”) into separate tokens, allowing for finer alignment and treebank annotation. Treebank annotation markup, such as “empty category” markers, is treated as separate tokens in the alignment annotation. Punctuation is also separated from preceding tokens.

With Chinese, segmentation is challenging due to the lack of word boundaries (Wu. [9]). Segmenting raw data into individual characters is the simplest kind of word segmentation, with each character being a token. More sophisticated segmentation schemes in MT systems group characters into words which consist of one or more characters. The word segmentation scheme proposed by the Penn Chinese treebank (CTB) team (Xue et al. [10]) is one of such schemes. We directly extract leaf tokens from the Penn CTB where the Penn CTB word segmentation scheme is applied. The extracted words are used for an intermediate alignment between character-level and larger syntactic unit alignments. To enforce data consistency and integrity, instead of segmenting raw files, we further segment the CTB-word segmentation files into character-based files, and thus following the *non-linear* approach. Each character and hyphen is a separate token, and punctuation is also separated from the preceding characters. The base-level alignment for our Chinese-English PAT begins at this character-level.

4. Alignment and Tagging Annotation

4.1 Levels of Alignment and Tagging

To build a PAT corpus, the data need to be aligned either at a specific level or at several levels. The base-level alignment is built on minimum translation units.

Upward, higher-level alignments are performed on larger linguistic units, such as tree-to-tree alignment. Generally, the base-level alignment is the *word* alignment. Arabic-English base-level alignment is at the word level. With Chinese, however, the minimum linguistic unit is a character. We chose the CTB for building the PAT, and the larger component alignment is the result of applying the CTB word segmentation scheme. Therefore, the alignment annotation at the LDC focuses on the Arabic-English word alignment, the Chinese character-level alignment, and the CTB word alignment. The first two are manual alignments while the CTB word alignment is automatically induced. To enrich the Chinese-English alignment, a layer of tagging annotation is performed manually on top of the character-level alignment and is automatically propagated to the CTB-word alignment.

4.2 Word Alignment Annotation

The task of word alignment is to identify correspondences between words, phrases or groups of words in a set of parallel texts. With reference to the Annotation Style Guide for the Blinker Project (Melamed, [5]), we developed two sets of alignment guidelines: Chinese-English and Arabic-English, which can be accessed from: http://projects.ldc.upenn.edu/gale/task_specifications/.

The guidelines discuss universal alignment approaches in addition to idiosyncrasies specific to the given language pair. General strategies and principles specify rules for annotating universal linguistic features, and specific rules are for idiosyncratic language features. The Arabic guidelines address Arabic-specific features, such as equational sentences, empty subjects, cliticization of determiners, prepositions, pronouns, and conjunctions, idioms and certain Arabic interrogative words with no equivalent words in English. For Chinese-English alignment, specific topics include the Chinese particles, non-inflection, topicalization, measure words, duplication, tense and aspects, various types of helping words.

Two types of links (*translated-correct* and *translated-incorrect*) and two types of markups (*not-translated correct* and *not-translated incorrect*) are designed to capture general linguistic information and language specific features. Most of the alignment links are *translated-correct* links which indicate valid translation pairs. *Translated incorrect* link type covers instances of erroneous translations lexically, grammatically or both. *Not-translated incorrect* refers to cases with a loss of semantic meaning and an absence of surface structure representation. For unaligned words, such as omissions or insertions of words, we use the *not-translated correct* markup to indicate cross-lingual features.

Two approaches are proposed for word alignment: *minimum match* and *attachment*. The *minimum match* approach, illustrated in Figure 3, aims to identify complete and minimal semantic translation units, i.e., atomic translation pairs. This method helps to map minimum syntactic structure unit equivalence, generating minimal semantic unit alignments which may be one-to-one, many-to-one or many-to-many links. The *attachment* approach is introduced to handle unaligned words.

The unaligned words are normally contextually or functionally required for semantic equivalence but they do not have surface structure translation equivalence. With the *attachment* method, shown in Figure 4, the unaligned words are attached to their constituent head words to indicate phrasal constituent dependency or collocation dependency. Unaligned words at the sentence or discourse level are not attached because they have no immediate constituents to depend on and attach to.

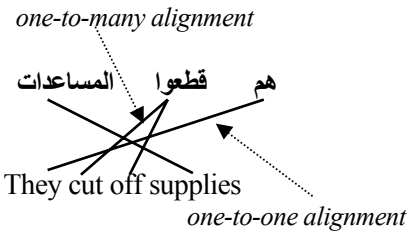

Figure 3: Minimum Match Approach

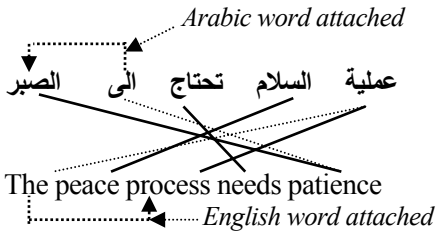

Figure 4: Attachment Approach

4.3 Tagging Annotation

To improve automatic word alignment and ultimately MT quality, researchers are exploring the possibility of incorporating extra information into word alignment. Following this direction, LDC collaborated with IBM in creating an additional layer of annotation by adding linguistic tags to the existing word alignments. Tags are added to both source and target languages to indicate different alignment types or functions of unaligned words. The tagging guidelines were jointly developed by LDC and IBM. The tags can be language independent, but the current tagging focus at LDC is the Chinese-English alignment. The Arabic alignment guidelines were updated to include a new word tag "GLU" for unaligned words, whereas for Chinese-English alignment, a set of tags were designed in the tagging guidelines for labeling all the aligned links and unaligned words (Li et al. [3]).

For Chinese-English alignment, we designed seven link types and fourteen word tags (Figures 5 and 6) to systematically address a variety of linguistic features.

Alignment Link Tags	***Examples***
Semantic	这个(this)***教授***(professor) [this ***professor***]
Function	***在***(in)这个(this)工厂(factory) [***in*** this factory]
Grammatically-inferred	***把工作***(work)完成(finish) [finish this ***work***]
Contextually-inferred	欢迎***收看CCTV***[Welcome to ***CCTV***]
DE-clause	离开(left)***的***女士(lady) [lady ***who*** has left]
DE-modifier	问题(issue)***的***(of)实质(nature) [the nature ***of*** this issue]
DE-possessive	教授(professors)***的***(from)关注(attention) [attention ***from*** the professors]

Figure 5: Link Types

Word Tags	*Examples*
Omni-function-preposition	***把***工作(work)完成(finish) [finish the work]
Tense/passive	暴露(exposed)***的***问题(issue) [the issue exposed]
Measure word	两(two)***家***杂志(magazines) [two magazines]
Clause marker	他(he)犯(made)错(mistake) [the mistake ***which*** he made]
Determiner	记者(reporter)说(said)… [***The*** reporter said…
TO-infinitive	继续(continue)工作(work) [continue ***to*** work]
Co-reference	主席(chairman)说(said)将要(would)…[The chairman said ***he*** would….]
Possessive	工厂(factory)工人(workers) [the workers ***of*** this factory]
DE-modifier	干(did)***得***快(fast) [did fast..]
Local context	欢迎(welcome)***收看***(CCTV) [Welcome to CCTV]
Rhetorical	台湾(Taiwan)学生(students)和(and)大陆(mainland) ***学生***(students) [students from mainland and Taiwan]
Sentence marker	老师(Teachers)很(very)忙(busy)***的***[Teachers are very busy.]
Context-obligatory	下雨(rains)了[***It*** rains]
Non-context-obligatory	他(He)***都***已经(already)离开(left)了[He already left]

Figure 6: Word Tags

The original alignment type *translated correct* is further classified into seven link types. The fourteen word tags are used for unaligned words. In the tagging guidelines, the Chinese 的 (DE) is a particular focus because of its complexities for machine translation (Li et al. [3]). To indicate the use of the particle 的 (DE), we tag all instances of this particle in Chinese texts by labeling them with DE-related alignment type and word tag, as illustrated with examples from Figures 5 and 6 above.

4.4 CTB Word Alignment and Tagging

The CTB word alignment is obtained from automatically transferring the manually-annotated character-level alignment. The transference merges the alignments if the CTB word has more than one Chinese character. We preserve the word tags for each individual character in this automatic alignment process. Similarly, link types are preserved to indicate the contextual information and different internal sub-part structures of CTB word alignment. Figures 7 and 8 illustrate how tags are preserved after automatic CTB word alignment. Figure 7 shows two aligned links at the character-level alignment. The Chinese token 1 (鲜) is aligned to the English token 2 (fresh), and the token 2 (花) is aligned to the tokens 1 and 3 (the flowers) (see alignment file format in Section 6). The link types are "semantic (SEM)" and "grammatically-inferred semantic (GIS)" respectively. The word tag DET is for "determiner". After the CTB word alignment processing (Figure 8), the CTB token 1 (鲜花) is aligned to the English tokens 1, 2, and 3 (the fresh flowers), and we keep both link types SEM and GIS to indicate contextual information.

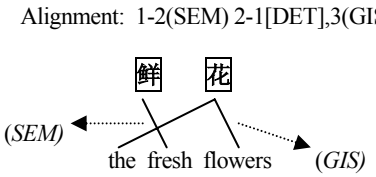

Figure 7: Character Alignment

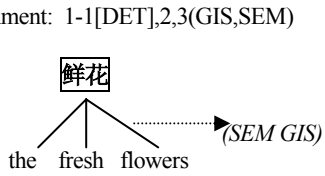

Figure 8: CTB-word Alignment

4.5 Efficiency and Consistency of Alignment and Tagging Annotation

To facilitate the annotation task, an annotation tool was developed at the LDC which allows alignment and tagging on the same interface. The annotation efficiency is monitored via the annotation workflow interface (Figure 9), where one can query the annotation volume and speed for a particular project, task, dataset, or annotator. The average annotation speed is about 8 hours per 10,000 source words for alignment and 6 hours per 10,000 source words for tagging.

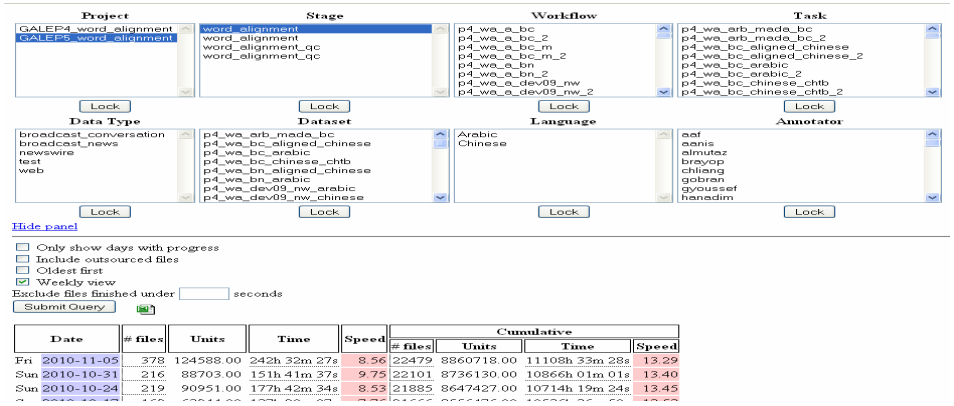

Figure 9: Efficiency Report Interface

To ensure annotation consistency, we conducted consistency tests on the pilot alignment of newswire data jointly annotated by LDC and IBM (Figure 10).

Data (Newswire)	*Chinese Characters*	*Precision*	*Recall*	*F-score*
File1	306	97.27%	95.70%	96.48%
File2	185	95.28%	96.19%	95.73%
File3	365	90.37%	91.20%	90.78%
File4	431	90.83%	92.61%	91.17%

Figure 10: Inter-annotator Agreement on Alignment

5 Treebank Annotation

Building PATs requires parallel treebanks. We use the Penn parallel treebanks for creating PATs at LDC. The Penn Arabic Treebank (ATB) annotation consists of two

phases: morphological/part-of-speech (POS) and syntactic/tree annotation. POS annotation includes morphological, morphosyntactic and gloss information. Syntactic annotation focuses on the constituent structures of word sequences, providing function categories for each non-terminal node, and identifying null elements, co-reference, traces, etc. (Maamouri et al. [4]). To build our Arabic-English PAT corpora, we started with treebank data from the most recent releases and ATB Part 3 (Bies et al. [1]). Treebank annotation markups are preserved during alignment process to maintain data integrity.

The Penn CTB corpora are segmented, POS tagged, and syntactically-annotated data. For our Chinese-English PAT corpora, we took all available CTB sources parallel to the English treebank for alignment annotation and corpora integration, excluding data with loose translations and files with improper format. The English translation treebank in correspondence to Arabic and Chinese is produced jointly by the Penn English Treebank team and the English treebank team at the LDC on four genres (BN, BC, NW and WB). For our Chinese-English and Arabic-English PAT corpora, we use English raw and tree files from the LDC published resources.

6 Data Structure and File Format

Instead of using .xml to construct the data, our PAT includes four text file types: raw, tokenized, word aligned, and treebanked data, one sentence per line without markups. Files with an identical filename base have the same number of lines, and the annotations of a specific line share the same line number. Data constructed this way is simple and straight-forward, keeping the integrity of annotation from each source while facilitating an easier annotation consistency check.

```
(TOP (IP (CODE 1) (NP-SBJ (PN 2)) (VP (ADVP (AD 3)) (VP (ADVP (AD 4)) (ADVP (AD
5)) (VP (VA 6)))) (PU 7)))
(TOP (IP (CODE 1) (NP-SBJ (PN 2)) (VP (ADVP (AD 3)) (ADVP (AD 4)) (ADVP (AD 5))
(VP (VV 6))) (PU 7)))
(TOP (IP (CODE 1) (NP-SBJ (PN 2)) (VP (ADVP (AD 3)) (ADVP (AD 4)) (VP (VV 5) (PU
6) (CP-OBJ-Q (IP (NP-SBJ (PN 7)) (VP (ADVP (AD 8)) (VP (VV 9) (VP (VV 10) (NP-
OBJ (NN 11) (NN 12)))))) (SP 13)))) (PU 14)))
```

Figure 11: Sample of Tree File

```
3-3(SEM) 4,5[MEA]-4[OMN],5(GIF) 6-6(TIN) -1[MET](MTA) -2[MET](MTA) 1[MET]-(MTA)
2[MET]-(MTA)
3-2(SEM) 6[CON]-(NTR) 7,8,9[DEM],10[LOC]-4[DET],5(COI) 2-3(FUN) 11,12-6(TIN) -
1[MET](MTA) 1[MET]-(MTA) 4[MET]-(MTA) 5[MET]-(MTA)
2,3-2(SEM) 7-4(TIN) 5,6-3(SEM) 4[COO]-(NTR) -1[MET](MTA) 1[MET]-(MTA)
11,12-8,9(SEM) 6[CON]-(NTR) 13,14-11(TIN) 4,5-3(FUN) 7-4,5,7(SEM) 2,3-2(FUN)
8[OMN],9,10-10(GIS) -1[MET](MTA) -6[MET](MTA) 1[MET]-(MTA)
```

Figure 12: Sample of Alignment File

The treebank and alignment files (Figures 11 and 12) do not contain token strings - only the token IDs which must be looked up in the tokenized file. Trees are represented in the Penn treebank format (labeled brackets). Tree leaves contain POS tags and token IDs corresponding to the numbers in the tokenized file. Most lines have one tree while some may have more. Multiple trees on one line are separated by whitespace. In a word alignment file, each line contains a set of alignments for a

given sentence, as shown in Figure 12, where the alignments are space-delimited, with each alignment in the format of "s-t(linktype)", *s* and *t* being a list of comma delimited source and translation token IDs respectively. The alignment type is in the parentheses and the word tags in square brackets.

7 Complications of Data Processing and Annotation

Integrating existing treebank annotation resources expedites the process of creating a PAT. However, as the down-stream annotation, the alignment process is challenging because of complications inherited from existing annotation resources.

The most common problem in data processing is segment mismatch. Mismatch may exist between source and translation raw files, between tree and raw files, and especially between translation tree and source language tree files. This problem arises when a single source sentence is translated into multiple independent English sentences. Treebank annotations of source and target all operate on single sentences. As a result, the number of source trees does not match that of target trees. We automatically re-align the mismatched sentences with an error rate below 5%. Errors resulting from this re-alignment are further handled during manual alignment annotation by rejecting the mismatched sentences. Other data processing complications include inconsistent filenames and file formats because the existing annotation resources involve different parties and various annotation stages. We standardized the filenames and converted the files into the desired release format.

Data from different sources create more noisy data for alignment annotation. Noisy data, the elements interfering normal annotation, refer here in the context of word alignment annotation to the sentences with incorrect translations/segmentations, sentences containing foreign language, or sentences that are ill-formatted. A "rejection" function is designed as a part of the alignment tool for annotators to reject such noisy data during annotation. Another type of noisy data is annotation markups carried over from up-stream annotation, for which a special tag is introduced.

8 Conclusion and Future Work

As an on-going project of the GALE (Global Autonomous Language Exploitation) program, this work has created large PAT corpora by aligning the existing parallel treebanks. Tagging annotation added to alignments is not the same as monolingual POS annotation, but rather helps to identify contextual and cross-lingual features which emerge in alignment process, thus contributing to alignment error reduction and high translation accuracy. Future efforts may scale up to richer tagging annotation, alignments of higher levels, and more language pairs.

Acknowledgement

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper

does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] Bies A., Mott J., Warner C. (2010). English Translation Treebank -- EATB Part 6 v2.0 (Annahar/ATB3 parallel). LDC Catalog Number: LDC2010E21.

[2] Ittycheriah, A. and Roukos, S. (2005). A Maximum Entropy World Aligner for Arabic-English Machine Translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 89-96.

[3] Li, X., Ge, N., Grimes, S., Strassel, S. M. and Maeda, K. (2010). Enriching word alignment with linguistic tags. In *Proceedings of LREC 2010*.

[4] Maamouri, M., Bies, A. and Kulick, S. (2008). Enhanced annotation and parsing of the Arabic Treebank. In *Proceedings of INFOS*.

[5] Melamed, D. (1998). *Annotation Style Guide for the Blinker Project* (URL: http://www.cs.nyu.edu/~melamed/ftp/papers/styleguide.ps.gz).

[6] Tiedemann, J., and Kotzé, G. (2009). Building a large machine-aligned parallel treebank. In *Proceedings of TLT'08*, pp.197–208, EDUCatt: Milano/Italy.

[7] Uchimoto, K., Zhang, Y., Sudo, K., Murata, M., Sekine, S. and Hitoshi, I. (2004). Multilingual aligned parallel treebank corpus reflecting contextual information and its applications. *In Proceedings of the Workshop on Multilingual Linguistic Resources*, pp. 63-70, Geneva: Switzerland.

[8] Volk, M., Gustafson-Capková, S., Lundborg, J., Marek, T., Samuelsson, Y. and Tidström, F. (2006). XML-based phrase alignment in parallel treebanks. *In Proceedings of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*, Trento, pp. 93–96.

[9] Wu, D. (2009). Toward Machine Translation with Statistics and Syntax and Semantics. IEEE Automatic Speech Recognition and Understanding Workshop*,* Merano: Italy.

[10] Xue, N., Xia, F., Chiou, F. and Palmer, M. (2005). The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 11(2):207-238.

Building Multilingual Named Entity Annotated Corpora exploiting Parallel Corpora

Maud Ehrmann, Marco Turchi

European Commission - Joint Research Centre (JRC), IPSC - GlobSec,
Via Fermi 2749, 21020 Ispra (VA) - Italy
E-mail: firstname.surname@jrc.ec.europa.eu

Abstract

This paper reports first experiments in the automatic building of multilingual named entity annotated corpora, taking advantage of a multiparallel corpus. We believe that providing such a resource could help to overcome the annotated data shortage in the Named Entity field and will guarantee comparability of named entity recognition system results across languages. Our approach is based on annotation projection, which is carried out with the help of a phrase-based statistical machine translation system. We obtain promising results and thus consider proceeding with other languages.

1 Introduction

Named Entity recognition is a well-established task: specified for the first time during the latest American MUC conferences, it is now acknowledged as a fundamental task to a wide variety of natural language processing (NLP) applications. Rule-based, machine learning and hybrid named entity recognition systems have been developed over the years, achieving respectable performances for various languages, domains and applications (Nadeau et al. [14]). As for many other NLP tasks, annotated corpora constitute a crucial and constant need for named entity recognition (NER). Within a development or training framework, annotated corpora are used as models from which machine learning systems (or computational linguists) can infer rules and decision criteria; within an evaluation framework, they are used as a gold standard to assess systems' performances and help to guide their quality improvement, *e.g.* via non-regression tests.

During the last decade, several named entity (NE) annotated corpora were built, thanks to a large series of evaluation campaigns (Fort et al. [7]). However, such resources remain rather rare and limited to a relatively small set of languages and domains. Even if unsupervised methods tried to overcome this difficulty, the shortage of annotated data for the large majority of world's languages remains a problem.

An obvious solution is to manually produce annotated corpora, but it is a complex and time-consuming task and it may be difficult to find experts in specific language.

Beyond annotated corpora's scarcity, another issue lies in the fact that annotation schemas or guidelines usually differ from one annotated corpus to another: named entity extents can be different (*e.g.* inclusion or not of the function in a person name, *Secretary of State Hillary Clinton* vs. *Hillary Clinton*), as well as entity types and granularity (*e.g.* some corpora may consider product names, whereas others will differentiate, within this category, vehicles, awards and documents, and others won't even consider product names). Such divergences should be expected, as annotated corpora are built within different frameworks and according to different applications. However, they constitute a real issue, particularly when developing or evaluating multilingual NE recognition systems. Actually, in a multilingual environment, if someone wants to use named entity annotated corpora (if available), he/she should first convert the data to a common annotation schema and document format before exploiting it. To avoid the annotation schema conversion step, Bering et al. [3] built a flexible evaluation tool; although efficient, this solution seems quite heavy to implement and requires a meticulous study of the different annotation schemas.

Our goal is to automatically build a set of multilingual named entity annotated corpora, taking advantage of the existence of parallel corpora (multiparallel or bilingual). Traditionally used in the field of Machine Translation, parallel corpora have been exploited in recent years in various NLP tasks, including linguistic annotation, with the creation of annotated corpora. The underlining principle is *annotation projection*, where annotations available for a text in one language can be projected, thanks to the alignment, to the corresponding text in another language, creating herewith a newly annotated corpus for a new language.

This method shows several advantages. Firstly it could be a way of overcoming NE annotated data shortage problem. Then, it could solve the non-harmonized annotation issue: if the projected annotations (on the target side) always come from the same automatic recognition system (on the source side), then we obtain annotated corpora in different languages, but with a common annotation schema. The use of multiparallel corpora also presents the benefit of ensuring the comparability of NER system results across languages; morever, as named entity recognition systems are domain-sensitive, it could be relevant to evaluate multilingual NER systems on equivalent tasks.

This paper relates our first attempt to apply this method to Named Entity annotations, projecting automatically annotated English entities to French, Spanish, German and Czech aligned corpora. Following this preliminary work, our objective is to automatically annotate and make freely available named entity corpora in a large set of languages, with a quality similar to that of manually annotated data.

The remainder of the paper is organized as follows. In section 2 we introduce related work; we then present our NE projection method (still at its first stage of development) in section 3, report the results in section 4 and finally conclude and propose some elements for future work in section 5.

2 Related Work

Regarding the automatic acquisition/building of NE annotated corpora, some work investigate how to constitute monolingual annotated data: An et al. [1] extract a huge amount of documents in Korean from the web and then annotate them automatically whereas Nothman et al. [15] make use of Wikipedia to create a named entity annotated corpus in English, transforming Wikipedia's links into NE annotations. In each case, the resulting corpora allow the authors to train a NER system that performs quite well, thus vouching for the newly labeled data quality.

With regard to parallel corpora, their exploitation has been growing in recent years, showing their usefulness in various NLP tasks like word sense disambiguation or cross-lingual tagging (refer to the state of art presented by Bentivogli et al. [2]). With respect to cross-lingual knowledge induction, multiple work addressed the challenge of automatic parallel treebank building, deducing syntactic information correspondences (Lavie et al. [12]) or projecting them from one language to another (Hwa et al. [8]). In addition, recent work carried out semantic information projection, mainly focusing on semantic roles and word senses (Padó et al. [16] and Bentivogli et al. [2]).

Several researchers investigated named entity annotation and parallel corpora exploitation. Klementiev et al. [9] proposed an algorithm for cross-lingual multiword NE discovery in a bilingual weakly temporally aligned corpus. Their goal is to extract pairs of named entities across languages, by co-ranking two clues: synchronicity (use of a time distribution metric) and phonetical similarity (use of a transliteration model). Ma [13] applies a co-training algorithm on unlabelled bilingual data (English-Chinese), showing that NE taggers can complement and improve each other while working together on parallel corpora. Samy et al. [17] developed a named entity recognizer for Arabic, leveraging an Arabic-Spanish parallel corpus aligned at sentence level and POS tagged. Yarowsky et al. [21] achieved some pioneer experiments, exploring the feasibility of annotation projection in four tasks, one of which was named entity annotation. The goal was to automatically induce stand-alone text analysis tools via robust annotation projection. Such approaches deal with named entity annotation and make use of parallel corpora but mainly aim at developing or improving NER systems; it seems that parallel annotated corpora are a positive side-effect of these work, but they don't go into details. Our approach differs in that we focus our attention on acquiring multilingual annotated corpora mainly for evaluation purpose. Therefore, high precision is required and we cannot afford noisy projections.

Finally, the work of Volk et al. [20] on combining parallel treebanks and geotagging offers similar results to what we propose, with the difference that they focus on the *location* type, ground the annotated entities with references to a gazetteer and work with a bilingual French-German corpus.

3 Named Entity Annotation projection

Given a multiparallel corpus and a monolingual NER system, our objective is to automatically provide NE annotations for each text of the aligned corpus. We assume that a possible solution to project a named entity from a text in one language to an aligned text in another language is to translate this entity, using different approaches, e.g. machine translation. Following this assumption, our multilingual NE annotation projection method relies, for the most part, on the use of a phrase-based statistical machine translation system (PBSMT). We used a multiparallel corpus in English, French, Spanish, German and Czech, that is news texts coming from the WMT shared tasks (Callison-Burch et al. [5]). For each language, we have a training set of roughly 70,000 sentence pairs and a test set of 2,490 sentence pairs. We used the test set for the annotation projection. The next sections detail each step of the NE annotation projection process.

3.1 Automatic annotation of source Named Entities

The first step is to annotate NEs in one corpus in a given language. We chose to annotate English entities of type *Person*, *Location* and *Organisation* and tried to project them in the corresponding texts in other languages. As a matter of fact English is a resource-rich language with already existing efficient tools, but one may choose another source language, according to his/her own goals and constraints.

We used an in-house NER system (Steinberger et al. [18] and Crawley [6]) to process the English source side text (any NER system or even manual annotation could have been used at this stage). It is obvious that the NER system quality is a crucial element that determines the projection quality: if the system misses one entity or wrongly annotates it, it won't be projected or it will be wrongly annotated. In our English text, the NER system annotated a total of 826 unique entities, corresponding to 1,395 entity occurrences, among them 649 person names, 412 location names and 332 organisation names.[1]

3.2 Source Named Entity translation

The second step corresponds to the translation of the previously extracted entities into French, Spanish, German and Czech. We firstly present the Phrase-Based Statistical Machine Translation system and account for its benefits in this particular task; we then report a correction phase and an evaluation of the NE translation.

Phrase-Based Statistical Machine Translation System. One of the most popular classes of statistical machine translation (SMT) systems is the Phrase Based Model [11]. It is an extension of the noisy channel model, introduced by [4], using phrases rather than words. A source sentence f is segmented into a sequence of I phrases $f^I = \{f_1, f_2, \ldots f_I\}$ and the same is done for the target sentence e,

[1] In this paper we do not go into details regarding the source NE annotation (type granularity, extents, etc.) as we focus more on the validation of the approach.

where the notion of phrase is not related to any grammatical assumption; a phrase is an n-gram. The best translation $\hat{e}$ of f is obtained by:

$$\hat{e} = arg\max_{e} p(e|f) = arg\max_{e} \prod_{i=1}^{I} \phi(f_i|e_i)^{\lambda_\phi} d(a_i - b_{i-1})^{\lambda_d} \prod_{i=1}^{|e|} lm(e_i|e_1 \dots e_{i-1})^{\lambda_{lm}}$$

where $\phi(f_i|e_i)$ is the probability of translating a phrase e_i into a phrase f_i. $d(a_i - b_{i-1})$ is the distance-based reordering model that drives the system to penalize significant reordering of words during translation, while still allowing some flexibility. In the reordering model, a_i denotes the start position of the source phrase that was translated into the ith target phrase, and b_{i-1} denotes the end position of the source phrase translated into the $(i-1)$th target phrase. $lm(e_i|e_1 \dots e_{i-1})$ is the language model probability that is based on the Markov chain assumption. It assigns a higher probability to fluent/grammatical sentences. λ_ϕ, λ_{lm} and λ_d are used to give a different weight to each element. For more details see [11].

Phrases and probabilities are estimated processing the parallel data. Word to word alignment is firstly extracted running the IBM models [4], and then, on top of it, proximity rules are applied to obtain phrases, see [11]. Probabilities are estimated counting the frequency of the phrases in the parallel corpus. In this work, we used the PBSMT system Moses [10].

Among all the possible translation techniques, we decided to use this approach because, in general, entities are a small set of contiguous words, phrases, and PBSMT systems perform better than systems based on single words. In this work, we do not apply the classical idea of translation: a sentence that is not present in the training data (unseen sentence) is translated to another language. In our experimental framework, we train a PBSMT system using as training data the parallel sentences that we want to annotate plus a larger set of parallel sentences. This means that the translation system knows how to translate the source entity, because it has seen it in the training data; this reduces the number of completely untranslated entities. At the end, we use the SMT system for its capability of aligning bilingual phrases across two parallel sentences more than for its capability of translating unseen sentences. Unfortunately, this experimental setting does not guarantee that all the source entities are always correctly translated, because its statistical approach favours those translations that appear more often in the training data. That's why we added a correction phase after the translation.

Correction phase. Entity translations are not always correct because the PBSMT system tries to reproduce the most readable sentence driven by the language model; in this way, the translation system may add articles, prepositions or in some cases groups of words before or after the entity name. For example, the french translation of *Afghanistan* is *en Afghanistan* and the translation of *Germany* is *l'Allemagne*. In these cases, only *Afghanistan* and *Germany* should be projected, as prepositions and articles cannot be part of proper names in French. We could observe similar phenomena in other languages.

To address this problem, we post-processed the translations in a simple way: applying stopword lists. This allowed us to correct a certain number of entities

for each language, even if some wrong entities could remain in the list. Before projecting these "corrected" translated entities in the aligned corpora, we asked bilingual annotators to check the correctness of the translated entities.

Evaluation of the NE translation. We randomly selected two hundred English entities and their relative translations in French, Spanish, German and Czech. We then provided annotators with the bilingual lists plus a set of evaluation categories that identify possible translation errors:

1. Correct Translation: the translated entity is correctly translated.
2. Extra Words: the translated entity contains some superfluous words (En: *tariq ramadan* Fr: *peut-être tariq ramadan*).
3. Missing Words: the translated entity does not contain some original words (En: *eastern punjab* Fr: *punjab*).
4. Wrongly Translated Words: the translated entity contains some words that are incorrectly translated (En: *reuters news agency* Fr: *nouvelle agence reuters*).
5. Wrong Word Order: some words in the translated entity are not correctly located (En: *south africa* Fr: *sud du afrique*).
6. Wrong Translation: the translated entity is wrong.

Evaluation results are reported in Table 1. In all languages, main problems seem to be the addition and subtraction of word(s) during the translation phase. This comes from the fact that the PBSMT tries to reproduce the most readable sentence (as pointed above), adding or removing some words that afterwards were not removed by the stoplists. We also observe that there are more completely wrong translations when French or Spanish are the target language. Presumably, this is due to the fact that there are different translation choices (verbatim or not) between languages for specific names such as *Canada Cup*, *Stanley Cup* or *Walmart Foundation*; in front of this situation, the annotators adopted different behaviours. We need to investigate this phenomenon, in order to know if we can predict when it is preferable not to translate, but to keep the English entity.

In general, SMT performance depends on the training set size [19]. We first trained the PBSMT system with the parallel sentences that we want to use in the projection only, obtaining poor results. For this reason, we then added more training data, whose size (70,000 sentence pairs) is still rather small according to the machine translation community standards. We believe that adding more data can increase the translation performance and in particular solve the problem of unwanted or deleted words in the translations.

	French	Spanish	German	Czech
Correctly translated	83,5	83.5	82.5	83.5
Extra words	4.0	3.0	7.0	9.0
Missing words	3.0	4.5	6.5	3.5
Wrong words	2.0	1.0	0.0	1.0
Wrong order	1.0	0.5	0.5	0.5
Wrong translation	6.5	7.5	3.5	2.5

Table 1: Human Evaluation of NE translation (error type percentages).

3.3 External Named Entity resource

In addition to the SMT approach, we benefit from an external multilingual named entity resource. The JRC's named entity database has been built up since 2004 through a daily analysis of tens of thousands of multilingual news articles per day; it contains, among others, translations and transliterations of entity names in several languages [18]. By querying this database, we retrieved, for each English entity, a list of translated entities (that may have different spellings) in a given language.[2]

The information coming from the external resource is quite reliable, because part of the entity names has been manually checked. However, it is not exhaustive. On the contrary, the SMT system provides translations almost every time, but they may be incorrect. In other words, information coming from the external resource and the SMT system can complement each other, the former boosting precision and the latter ensuring recall. For example, *Sakharov Prize for Freedom of Thought* is correctly translated by the SMT system for each language while the database does not contain this name.

3.4 Named Entity projection

Once we have a list of possible translations for a given NE in a particular sentence, we try to project it into the corresponding sentences of the aligned corpora, using a simple and strict string matching: the translation is present or not. We applied the whole processing chain to our multiparallel corpus; the next section presents the projection results.

4 Results and discussion

We evaluate the performance of the projection using three different translation approaches. English entities are translated using: (1) external information: for each

[2]The database contains 134,046 en-fr named entity translations, 157,442 en-sp, 156,363 en-de and 2,807 en-cs.

language pair, a list of English-Foreign entity associations is used as a look-up table (*Ext* in Table 2), (2) machine translation system (*SMT*) and (3) external information and machine translation system together: a list of all possible translations is associated to each English entity[3](*All*).

As we do not have a reference corpus, we can only compute projection's Recall. An indirect way to evaluate the Precision is the SMT evaluation, but this is only a partial evaluation. In the future, we will vary our projection method (not only strict string matching) and manually annotate a part of the multilingual set to provide a complete evaluation of the projection.

During the first step (source NE annotation), we noticed the presence of wrong English entities. In this work, we do not evaluate the quality of the NER system that we used, but we are interested in evaluating how it affects our projection performance. For this purpose, we manually corrected the English entities. In Table 2, we report results for projections done using all the English entities and only the correct ones. Recall is computed relative to the total number of English entities.

	French	Spanish	German	Czech
Ext	0.325	0.264	0.291	0.103
Ext (En Correct)	0.343	0.278	0.306	0.106
SMT	0.798	0.787	0.794	0.535
SMT (En Correct)	0.825	0.806	0.813	0.545
All	0.807	0.800	0.807	0.547
All (En Correct)	0.834	0.821	0.827	0.557

Table 2: Recall of the annotation projection.

The first observation is that projections are strongly affected by the target language. When French, Spanish and German are the target languages, performances are similar, while with Czech there is a drastic drop in performance. This is due to the fact that Czech is a highly inflected language and for the same English entity there are more than one possible translation (morphological variants).

Projections obtained using only the external resource produce low recall. This approach is quite good for those English entities that have a standard form like first name-surname (e.g. *Matt Damon*) or location names (e.g. *South Africa*), but is less efficient for organization entities (e.g. *Czech hydrometeorological institute*). The big advantage of using an SMT system trained with the data that we want to use during the projection is that all the information is available for the SMT system which can correctly translate entities, even complex ones. This aspect can be seen in the results, where recall with SMT translation improves substantially compared to the recall obtained using the external resource only. Merging of external and SMT translations produces small improvements, while removal of wrong English entities affects positively the results, in particular for German, Spanish and French.

[3]If more than one translation matches the target sentence, it is counted only one time.

5 Conclusion and Future Work

Parallel corpora can support the automatic creation of multilingual NE annotated corpora. We presented preliminary experiments of a NE annotation projection method for a 5 language multiparallel corpus, obtaining encouraging results.

The current approach can be improved in several ways. First of all, as demonstrated by different results with/without wrong English entities, we need to improve the NER system. Then the projection approach (presence/absence of the translated entity) is particularly strict. We believe that different methods based on word similarity and word alignment can be used to find the correct entity in the target sentence.The main issue is the projection of the entities in a highly inflected language. To solve this problem, one solution is to force the PBSMT system to emit also the less probable translations, trying to cover all possible variations in the inflected language. Finally, we intend to apply this method to other parallel corpora in different languages.

Acknowledgements We would like to thank Josef Steinberger and Ralf Steinberger for accepting to annotate Czech and german entities, as well as J. Brett Crawley, Jenya Belyaeva, Vanni Zavarella and again Ralf for their comments.

References

[1] An, J., Lee, S. and Lee, G. (2003) Automatic acquisition of named entity tagged corpus from world wide web. In *Proceedings of the 41st Annual Meeting on ACL (ACL'03)*, Sapporo.

[2] Bentivogli, L. and Pianta, E. (2005) Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. In *Natural Language Engineering* pp. 247–261, Cambridge University Press.

[3] Bering, C., Drozdzynski, W., Erbach, G., Guasch, C., Homola and others. (2003) Corpora and evaluation tools for multilingual NE grammar development. In *Proceedings of Multilingual Corpora - Linguistic Requirements and Technical Perspectives*, Lancaster.

[4] Brown, P.F., Della Pietra, S., Della Pietra, V.J. and Mercer R.L.(1994). The Mathematic of Statistical Machine Translation: Parameter Estimation. In *Computational Linguistics*, 19(2):263–311.

[5] Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J. (2009) Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth WMT'09*, Athens.

[6] Crawley, J. B. and Wagner, G. (2010). Desktop text mining for law enforcement. In *Proceedings of ISI'10*, Vancouver.

[7] Fort, K., Ehrmann M. and Nazarenko, A. (2009) Towards a Methodology for Named Entities Annotation. In *Proceedings of LAWIII*, Singapore.

[8] Hwa, R., Resnik, P., *et al*. (2005). Bootstrapping parsers via syntactic projection across parallel texts. In *Natural Language Engineering*, 11(3).

[9] Klementiev, A. and Roth, D. Named Entity Transliteration and Discovery from Multilingual Corpora. (2008) In *Learning Machine Translation.* MIT Press.

[10] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N. and others (2007). Moses: Open source toolkit for statistical machine translation. ACL, 45(2), Columbus, Oh, USA.

[11] Koehn, P. (2010). Statistical Machine Translation. Cambridge Univ. Press.

[12] Lavie, A., Parlikar, A. and Ambati, V. (2008). Syntax-driven learning of subsentential translation equivalents and translation rules from parsed parallel corpora. In *Proceedings of theHLT-SSST-2 workshop*, Columbus, Ohio.

[13] Ma, X. (2010) Toward a Name Entity Aligned Bilingual Corpus. In *Proceedings of the Seventh LREC Conference*, Malta.

[14] Nadeau, D., and Sekine, S. (2007) A survey of named entity recognition and classification. In *Linguisticae Investigaciones*, 30-1, pp. 3-26.

[15] Nothman, J., Curran, J., and Murphy, T. (2008) Transforming Wikipedia into named entity training data. In *Proceedings of the ALTA Workshop*, Hobart.

[16] Padó, S. and Lapata, M. (2009) Cross-linguistic projection of role-semantic information. In *Journal of Artificial Intelligence Research*, 36.

[17] Samy, D., Moreno-Sandoval, A. and Guirao, J.M. (2005). A Proposal for an Arabic Named Entity Tagger Leveraging a Parallel Corpus (Spanish-Arabic). In *Proceedings of RANLP Conference*, Borovets, Bulgaria.

[18] Cross-lingual Named Entity Recognition. Steinberger, R. and Pouliquen B. (2007). In *Linguisticae Investigationes*, 30-1, John Benjamins.

[19] Turchi, M., DeBie, T. and Cristianini N. (2008). Learning Performance of a Machine Translation System: a Statistical and Computational Analysis. In *Proceedings of the Third WMT'08*, Columbus, Oh, USA.

[20] Volk, M., Goehring, A. and Marek, T. (2010) Combining Parallel Treebanks and Geo-Tagging. In *Proceedings of The Fourth LAW Workshop*, Uppsala.

[21] Yarowsky, D., Ngai, G. and Wicentowski, R. (2001) Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT'01*, San Diego.

Clause restructuring in English-Swedish translation

Lars Ahrenberg

Department of Computer and Information Science
Linköping University
E-mail: lars.ahrenberg@liu.se

Abstract

Medium rank clauses, such as participial and infinitival clauses, have been shown in earlier studies to be more frequent in English than in Swedish. Instead Swedish prefers complete, finite clauses. This constitutes a problem for English-Swedish machine translation. Here I report a study of such constructions using the LinES Parallel Treebank. I also show how the dependency annotation in LinES can be used to define clauses of different ranks.

1 Introduction

Clause structure is a major aspect of syntax and also one in which languages differ. Mastering clause structure means not only being able to produce grammatically well-formed clauses, but also being able to select the right type of clause in the right context. This is of special importance in translation, where source language norms may clash with target language norms. A good human translator would have developed a good sense of linguistic differences in this respect, but machine translation systems are often vulnerable to the influence of source language clause structure.

In statistical machine translation the type of restructuring that has been considered most is reordering. Several studies have shown improved performance when clause constituents are reordered to meet the norms of the target language e.g., [5], [6]. However, reordering is not the only relevant aspect of restructuring; additions and deletions of major constituents may be necessary or preferred as well as shifts of verbal morphology. In this paper our focus is on tenseless and subjectless constructions, which tend to be more common in English than in Swedish. The following are two examples where we compare a human translation to the translation suggested by Google Translate[1]:

[1]The first example is a variant of an authentic example from the LinES corpus, with translation by the author, the second example is taken from the Harry Potter section of LinES.

(1) EN: When creating a copy, she uses a very sharp point.
SE: När hon gör en kopia använder hon en mycket vass udd.
Gloss: When she creates a copy, she uses ...
Google: När du skapar en kopia, använder hon en mycket vass spets.
Gloss: When you create a copy, she uses ...

(2) EN: She leered at him, showing mossy teeth.
SE: Hon gav honom ett illvilligt leende, som avslöjade maskstungna tänder.
Gloss: She gave him a malicious smile that revealed mossy teeth.
Google: Hon sneglade på honom, visar mossiga tänder.
Gloss: She leered at him, shows mossy teeth.

For both examples Swedish prefers a finite clause construction with an overt subject or subject place-holder. This means generating both a tense and a subject that are congruent with the context. Google Translate manages to generate a tense, and, in sentence (1) also a subject, but does not succeed in enforcing the requirements on congruence. While other translations are possible here, literal translations using a Swedish participial form are not.

In linguistically oriented translation studies changes of this kind have been studied and classified by many authors, e.g., as category shifts [4], or transpositions [11]. Since they introduce material which is only implicit in the source, they may also be regarded as explicitations. Here I will call them shifts of clause rank, or simply rank shifts, after [7]. In this work Rune Ingo compares Finnish and Swedish on the one hand, and English and Swedish on the other. One claim of his is that participial and infinitival clauses are more common in Finnish and English than they are in Swedish, supporting the claim with percentages from different kinds of corpora. He also argues for the position that, normally, the translator should produce different types of clauses in the proportions that are suitable for the target language and the given text genre.

In this paper I treat Ingo's claim as an hypothesis to be tested for English source texts and Swedish translations, using the LinES English-Swedish parallel treebank [1] as test data. The hypothesis, then, is that the English half of the corpus contains more instances of participial and infinitival clauses than the Swedish side, and more specifically, that we will find a significant number of cases, where such clauses have been translated by clauses of a higher rank, in particular finite clauses. I will also investigate to what extent there are differences among the different text types included in LinES.

This, in turn, raises the question whether different clause types can be recognized with high accuracy in a corpus where the syntactic annotation does not explicitly mark them. Thus, another aim of the paper is to provide definitions of various clause types using the annotation in the treebank. The study as a whole can be taken as a support for the view that syntactically annotated parallel corpora are useful for translation studies. While parallel corpora have been recognized as

primary sources of data in many areas including translation studies and translation training ([3], [9]) they are not usually annotated syntactically. However, the range of linguistic and translational phenomena one can study is very much dependent on the available corpus annotation.

The rest of the paper is organized as follows. In the following section I introduce Ingo's model of clause ranks [7]. Section 3 presents our data and the annotation used. Section 4 explains the method and the use of the annotation for the purpose of the study. Section 5 presents our findings, and Section 6, finally, states the conclusions.

2 Clause types and clause ranks

Ingo's model of clause ranks has six levels:[2]

major clause,	e.g. *they arrived in London; Kim plays the guitar*
minor clause,	e.g. *when they arrived in London; that Kim plays the guitar*
participial clause,	e.g. *arriving in London; (while) playing the guitar*
infinitival clause,	e.g. *(them) to arrive in London; (ask her) to play the guitar*
nominalization,	e.g. *their arrival in London; Kim's guitar play*
without predication,	e.g. *in London, they... ; on guitar, Kim*

The further down we go in this hierarchy, the more the constructions lack features of a complete clause. These features are, according to Ingo, (i) the presence of a subject, (ii) the presence of a tense marker, (iii) the marking of mode, (iv) the optional presence of a negation.

3 The data

The parallel treebank used for this study comprises four subcorpora as outlined in Table 1: on-line help texts for MS Access for Windows XP (Access), Europarl data, and excerpts from two novels. Each subcorpus used for the study has a size of roughly 600 sentence pairs. The syntactic annotation employs parts-of-speech, morphological properties, and dependency functions. Every sentence is assumed to have a unique head, marked by the function 'main', and all other tokens, except punctuation marks, are direct or indirect dependents of the head. Monolingual files are XML-formatted. An annotated segment pair is shown in Figure 1.

The dependency annotation employed in LinES is surface-oriented and projective, making it easy to convert into a phrase-structure tree. The monolingual files were first parsed by Connexor parsers for English and Swedish [10] but the actual

[2]The English terms are translations from the Swedish ones used by Ingo: *clause rank: satsgrad; major clause: huvudsats; minor clause: bisats; participial: particip; infinitive: infinitiv; nominalization: nominalisering; without predicate: predikatslös.*

<s id="s3">
<w .. relpos="1" base="noone" func="subj" fa="2" pos="PRON" msd="SG">Noone< /w>
<w .. relpos="2" base="be" func="main" fa="0" pos="V" msd="PRES">is< /w>
<w .. relpos="3" base="very" func="ad" fa="4" pos="ADV">very< /w>
<w .. relpos="4" base="patient" func="sc" fa="2" pos="A" msd="ABS">patient< /w>
<w .. relpos="5" base="." pos="FE" msd="Period">.< /w>
< /s>

Figure 1: Morphosyntactic annotation of an English sentence in LinES.

annotation employs a different set of values, and for some constructions, different analyses. All annotations, including dependencies and alignments have been manually reviewed.

Subcorpus	**Text type**	**Sentences**	**Src words**	**Trg words**	**Ratio**
Access	online help texts	595	10,451	8,898	0.85
Europarl	political debates	594	9,334	8,715	0.93
Bellow[3]	fiction	604	10,310	9,962	0.97
HarryP[4]	fiction	600	10,171	10,501	1.03
Sums:		2393	40,266	38,076	0.95

Table 1: Corpus overview showing text type and size.

The word alignment is based both on semantic and structural correspondence where many-to-many alignments (as usual) represent corresponding units that cannot be analysed into smaller (1-1, 1-n, or n-1) alignments. Alignment was performed interactively using the I*Link tool [8]. Word alignments are complete, i.e., a decision has been made for each token in the corpus if, and how, it corresponds to something in the other language. A word link is represented as a paired list of indices such as (4-5/1) which says that the 4th and 5th words of the source sentence have been linked to the first word of the target sentence. The alignment encoding for the sentence in Figure 1 and its Swedish translation is shown in Figure 2.

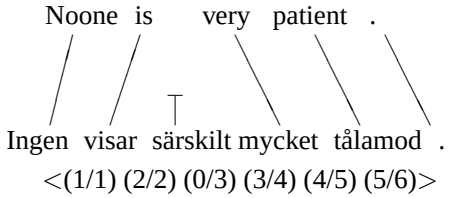

Figure 2: Swedish translation and alignment for the English sentence in Figure 1.

Null links are represented by the number 0. For example, (0/3) means that the third word of the Swedish sentence is judged to have no correspondent in the

English sentence.

4 Defining clause ranks

Our basic approach is to identify all clauses in the corpus, and then classify them in terms of a given clause rank. We restrict this process to clauses that are governed by verbs or participles, i.e., to words that have the part-of-speech annotation pos="V" or pos="PCP". Since each clause has a single governor we can identify corresponding pairs of clauses through the word alignment. If the translation is not a clause, but a phrase of some sort, we can still identify the image and its properties.

4.1 LinES annotation of clause elements

The notion of clause is underlying the LinES annotation, since a subset of the dependency relations are defined to apply only at clause level, i.e., to relate words to a verbal item. Similarly, other dependencies are restricted to noun phrases, while others, such as the coordination dependency can have almost any type of governor. The clause-level dependency relations used in LinES are listed in Table 2.

Distinctions between different clause types, however, are not part of the annotation scheme, and cannot be seen in the definitions of categories or dependency relations. In particular, Ingo's system of clause ranks formed no part in its development. It is therefore something of a test for the LinES scheme to model clause ranks within the scheme.

Label	Explanation
vch	Auxiliary verb or infinitive marker
advl	Adverbial
subm	Subjunction
subj	Subject
obj	Object (direct or indirect)
sc	Subject complement
oc	Object complement
prt	Particle
pobj	Oblique object
initm	Initiator e.g. an interjection
ad	General pre-modifier
cc	Coordinating conjunction or conjunct

Table 2: Clause-level dependency relations in LinES annotation. Clause-specific relations above the horizontal line.

4.2 Clause rank definitions

For the definition of clause ranks we need to consider the properties of the clause governors and the relations to their dependents. In some cases, such as the property of being tensed, the subtree dominated by a clause governor needs to be searched into sufficient depth. As in [10] auxiliaries, including the infinitive markers, are related to their governors via the dependency 'verbal chain element' (vch). Usually the tense marker will appear on the first element of the verbal chain, and the whole chain needs to be searched. Also the subject is a dependent of the first element of the verbal chain, rather than the main verb, when there are auxiliary verbs.

Clause rank	**Features**
major clause	+Tensed, +Subject, +Main
minor clause	+Tensed, +Subject or +Subjunction, -Main,
coordinated VP	+Tensed, -Subject, +ccVerb
participial clause	-Tensed, -Subject, +Participial, -Attributive
participial attribute	-Tensed, -Subject, +Participial, +Attributive
infinitival clause	-Tensed, -Subject, +Infinitive

Table 3: Clause rank feature analysis.

The main features that distinguish the different clause ranks are the following:

+Tensed, the presence of a tense marker on the first element of the verbal chain.

+Participial, the chain is -Tensed and the first element is a past or present participle.

+Infinitive, the chain is -Tensed and the first element is the infinitive marker or a verb in infinitive form.

+Subject, the presence of a word contracting the subject relation to the verbal chain. Imperative verbs usually don't have explicit subjects, but are assigned this feature by default.

+Subjunction, The presence of a subjunction or phrase having the 'subm' relation to the verbal chain.

+Main. the property of being the governor of an entire segment. The opposite, -Main, means being governed. However, a clause that expresses direct speech is also categorized as +Main in this study, while being annotated as an object of a communicative verb.

+ccVerb, the property of being a conjunct of a main verbal item, possibly through a chain of coordinations.

+Attributive, the property of being an attribute of a noun. This implies being +Participial.

The clause ranks are defined as conjunctions of these features. The definitions of the clause ranks are summarized in Table 3.

From Ingo's description it is not clear how conjoined clauses should be treated. I have chosen to define complete clauses that are coordinated with a main clause as major, whereas verb phrases that lack an explicit subject are given a category of their own even though they may be coordinated with a main (or subordinate) clause.

There are a few problems in the above definitions with respect to how well they capture the intended concepts. One concerns the distinction between participles on the one hand, and adjectives and nouns on the other. This distinction can be drawn in different ways, but the category participle tends to be a bit overused in LinES, as the basic criterion is a formal one. While participles are more common in English, this tendency is the same for both English and Swedish. Another decision that has effects on the numbers is that verbs with similar meanings may be classified as an auxiliary for one language, but as a non-auxiliary for the other. Also, annotation errors can still be found that affect the classification.

5 Results

Using the clause rank definitions we can simply count the number of clauses at each rank. For the reasons given in the previous section, the figures reported, see Table 4, are not exact but are stable enough to reflect tendencies. They give support to our hypotheses with one exception. In agreement with the hypotheses, Swedish has more finite clauses and, in particular, more minor clauses than English. The number of participial clauses on the English side is more than four times as many as on the Swedish side. But, contrary to the hypothesis, the Swedish side also has more infinitival clauses than the English side.

Table 4 also shows that the tendencies are quite stable across the sub-corpora. The Europarl corpus is slightly divergent, with fewer major clauses on the Swedish side than the English side and almost equal number of infinitival clauses. This is not surprising given that it partly contains parallel translations and has been shown to differ from the other three subcorpora in other respects as well [2].

Clause rank	**Access**		**Europarl**		**Bellow**		**HarryP**		**Sums**	
	En	**Se**	**En**	**Se**	**En**	**Se**	**En**	**Se**	**En**	**Se**
major clause	473	492	598	572	640	655	676	700	2387	2419
minor clause	298	403	261	327	289	371	305	417	1153	1518
coordinated VPs	33	41	14	26	39	72	91	159	177	298
participial clause	273	37	131	37	150	52	217	56	771	182
infinitival clause	210	243	176	172	149	200	138	189	673	804

Table 4: Frequency of clauses at different ranks distributed on sub-corpora.

To see how the different ranks have been treated in translation, we need to exploit the alignment. As the alignment is based on words, it may happen that single words are aligned to word sequences on the other side. Nevertheless, we

Source side clause ranks	**Target side clause ranks**						
	Major	Minor	ccVP	Pcpcl	Infcl	NP	Other
participial clause	78	184	81	56	175	50	149
infinitival clause	83	110	14	1	404	16	45

Table 5: Frequencies of mappings for English clauses of medium rank.

Clause rank with function	Major or minor	ccVP	Pcpcl	Infcl	Other
Adverbial participial clause	51	35	17	10	41
Nominal participial clause	52	5	0	111	48
Modifying participial clause	103	5	17	6	60
Adverbial infinitival clause	38	4	0	66	9
Nominal infinitival clause	90	3	0	137	25
Modifying infinitival clause	39	0	1	79	13

Table 6: Frequencies of mappings depending on grammatical function.

can look at clause governors and their individual images and see to what extent that image includes a clause governor on the other side, and, if so, what rank that clause belongs to. It may also happen that a clause is nominalized in translation, and we include those cases as well. However, the fact that a verb or participle is aligned with a noun does not guarantee that the image is a nominalization; it may, for example, be the result of a single verb being mapped to a complex verb construction such as English *decide* being translated by Swedish *fatta beslutet* ('make the decision').

We focus on participial and infinitival clauses as these are the ones showing the greatest differences in numbers. Data for these two ranks are shown in Table 5. We can see that English participial clauses, when translated into Swedish, yield both clauses of higher rank, and clauses of lower rank. The most common translations are minor clauses and infinitival clauses, while only about 6% of the translations are Swedish participial clauses. The category 'Other' also has many instances, the majority being distributed on prepositional and adjectival phrases, and deletions. Infinitival clauses are sometimes translated into higher ranks, but a large majority, 60%, are translated as infinitival clauses.

The function of the clause has an impact on restructuring. If we divide the participial and infinitival clauses into different groups depending on whether they have an adverbial, modifying or nominal (subject, object, or oblique object) function, we can see that for participial clauses, adverbial and modifying clauses tend to be rendered as complete clauses to a much larger extent than those with nominal functions, as shown in Table 6. For infinitival clauses the function has less impact on the proportion of cases that are rendered as complete clauses.

It can be seen that about 40% of the participial clauses are translated by tensed

clauses or phrases, and that almost 30% also have an overt subject. For English infinitival clauses the corresponding proportions are just below 30%. Even if the numbers for participial clauses on the English side may be exaggerated there are a significant number of instances where human translators select a higher clause rank than the one appearing in the English source and thus, supplying a tense, and a subject or a place-holding subordinator that are congruent with the context. Some examples are given in Table 7.

Mapping	**Clause pair**
pcpcl → minor (Access)	EN: In MS Access 2000 *using ADOX* SE: *När du använder ADOX* i MS Access 2000
pcpcl → major (Bellow)	EN: *Anticipating a difficulty*, I ask the stewardess... SE: *Jag förutser ett problem* och ber flygvärdinnan...
pcpcl → major (Europarl)	EN: ... is also imprudent in *introducing issues not included...* SE: ... *tar* också på ett oförsiktigt sätt *upp frågor som inte förekom...*
pcpcl → minor (Harry P)	EN: ... felt right into the corners *before sweeping the whole lot...* SE: ... kände efter långt inne i hörnen *innan hon sopade ner allt...*
pcpcl → major (Access)	EN: A different layout lets you *calculate and compare...* SE: Med en annan layout *kan du beräkna och jämföra...*
infcl → minor (Harry P)	EN: ..punish himself most grievously *for coming to see you* SE: ..bestraffa sig själv ytterst hårt *för att han hälsat på*

Table 7: Examples of high rank Swedish translations of medium rank English clauses.

6 Conclusions

A dependency-based annotation scheme which notionally distinguishes relations at the clause level and relations at the phrase level can also be used to identify clauses of different ranks. This allows hypotheses as regards restructuring at the clause level in translation to be tested and instances of such changes to be investigated in more detail, whether by human or machine translators.

The LinES data largely confirms the hypothesis that clauses without tense and subjects are more common in English than in Swedish translations. However, in LinES, this is entirely due to participial clauses, while the infinitival clauses are more common on the Swedish side than on the English side. Still, in a sizeable number of cases human translators selects a clause type of higher rank, with material that needs to be congruent with the context. This would seem to pose a problem for current approaches to statistical MT.

References

[1] Ahrenberg, Lars (2007) LinES: An English-Swedish Parallel Treebank. In *Proceedings of The 16th Nordic Conference of Computational Linguistics*, Tartu, Estonia.

[2] Ahrenberg, Lars (2010) Alignment-based profiling of Europarl data in an English-Swedish parallel corpus. Proceedings of the Sixth Conference on Language Resources and Evaluation (LREC'2010), Malta, 19-21 May, 2010.

[3] Baker, Mona (1993) Corpus Linguistics and Translation Studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli (eds.) *Text and Technology*, Amsterdam and Philadelphia: John Benjamins, pp. 233-250.

[4] Catford, J. C. (1965) *A Linguistic Theory of Translation: An Essay in Applied Linguistics.* London, Oxford University Press.

[5] Collins, Michael, Koehn, Philipp, and Kucerova, Ivana (2004) Clause restructuring for statistical machine translation. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, pp. 531 - 540.

[6] Elming, Jakob (2008) Syntactic reordering integrated with phrase-based SMT. In *Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation (ACL-08 SSST-2)*. Columbus, Ohio, USA.

[7] Ingo, Rune (2007) *Konsten att översätta: översättningens praktik och didaktik*. Lund, Studentlitteratur.

[8] Merkel, Magnus and Michael Petterstedt and Lars Ahrenberg (2003) Interactive Word Alignment for Corpus Linguistics. Proceedings of Corpus Linguistics 2003. UCREL Technical Paper No 16.

[9] Saldanha, Gabriela (2009) Principles of corpus linguistics and their application to translation studies research. *Revista Tradumàtica - Traducció i Tecnologies de la Informació i la Comunicació* No. 07.

[10] Tapanainen, Pasi and Timo Järvinen (1997) A non-projective dependency parser. In Proceedings of the 5th Conference on Applied Natural Language Processing, Washington, D.C, April 2007, pp. 64-71, Association for Computational Linguistics.

[11] Vinay, J-P and J. Darbelnet (1977) *Stylistique comparée du français et de l'anglais.* Paris, Didier.

Towards Parallel Czech-Russian Dependency Treebank

Natalia Klyueva and David Mareček

Charles University in Prague
Institute of Formal and Applied Linguistics,
E-mail: {kljueva, marecek}@ufal.mff.cuni.cz

Abstract

In this paper we describe initial steps in constructing a Czech-Russian dependency treebank and discuss the perspectives of its development. Following the experience of the Czech-English Parallel Treebank we have taken a syntactically annotated "gold standard" text for one language (Russian) and run an automatic annotation on the respective parallel text for the other language (Czech). Our treebank includes also automatic word-alignment.

1 Introduction

Large number of treebanks has appeared recently, and constructing the parallel treebanks is becoming more popular. This type of linguistic data presents valuable resource for both theoretical research in comparative syntax and NLP applications like Machine Translation. Parallel treebanks are generally compiled for English and some other language, but exceptions exist. To the best of our knowledge, no such parallel treebank exists for related Slavic languages.

We have created a small parallel treebank using data and tools from two existing treebanks. The manually annotated Russian data are taken from SynTagRus treebank [8]. Tools for the parsing the corresponding text in Czech are taken from the TectoMT framework [10]. We believe that our parallel treebank will open a road to the development of such treebanks for other Slavic languages.

Our project is connected to PCEDT - the Prague Czech-English Dependency Treebank [2]. Data for an annotation in it were taken from the Penn Treebank, precisely, a part which contains the texts from the Wall Street Journal. Though our project can not be compared to PCEDT in both quality and quantity, as the translation from English into Czech was made as closely to the original as possible, and the size of it is suitable for NLP tasks, for example Machine Translation.

As in PCEDT, we borrowed the text annotated within another framework and transformed it into a PDT style. It was easier for us because both treebanks anno-

tate dependency structure, not phrase structure. Still, we were not able to manually check the automatically parsed Czech text as it is done in the PCEDT.

Another very similar project is SMULTRON [1], the English-German-Swedish multilingual treebank, which also disposes a set of tools, as for example the Tree Aligner, that may be useful for our Treebank in the future.

This paper is structured as follows. Section 2 provides an overview on the two treebanks - the SynTagRus for Russian and the PDT for Czech, here we also introduce the data we chose for our treebank. In Section 3 an adaptation of the Russian annotation schema to the PDT style is described. Section 4 demonstrates the process of an automatic annotation of Czech text. Section 5 overviews the core – compilation and description of the treebank. Section 6 provides an example of a treebank exploitation. Finally, we conclude in Section 7.

2 Data and Tools

2.1 PDT and TectoMT

We decided to choose the Prague Dependency Treebank as a platform for our treebank, as it is more experienced with a parallel treebank handling and dispose tools for this. PDT contains 115,844 sentences from newspapers and journals.

In Prague Treebanking school a sentence is annotated on three layers: morphological, analytical and, tectogrammatical.

2.1.1 The Morphological Layer

Each word in a tree is represented as a node with a lemma and a tag assigned. The morphological tag is so-called positional, 15 positions are filled with an appropriate morphological category (Part of Speech, Gender, Number, Case, Person, Tense, etc.). All the sentences in PDT are annotated on this level.

2.1.2 The Analytical Layer

Syntactic annotation is presented in form of dependency tree, where each morphologically annotated token from the previous level becomes a node with an assigned analytical function. Analytical function (afun) reflects a syntactic relation between a parent and a child node and is stored as an attribute of the child. Examples of an analytical functions: Subject (Sub), Predicate (Pred), Object (Obj) etc. Analytical layer is annotated in 75 % of PDT texts.

2.1.3 The Tectogrammatical Layer

The annotation on the tectogrammatical layer (t-layer) goes deeper towards the level of meaning. Function words (prepositions, auxiliary verbs, conjunctions, etc.) are removed from the correspondent analytical tree and are stored as attributes

of autosemantic words, leaving only content words as the nodes on the t-layer. Tectogrammatical layer is annotated on 45 % of PDT texts.

The tools for automatic annotation of Czech sentences on these three layers are freely available in a TectoMT framework [10], which is used in our work too.

2.2 SynTagRus

SynTagRus is a collection of texts annotated on a morphological and a deep syntactic level. Texts in SynTagRus are mainly newspaper articles with a small amount of modern prose texts, it contains approximately 460,000 words. The treebank is coded in an XML-based schema.

Words are represented by nodes, which have three morphological attributes: word form, lemma and tag. Unlike a Czech positional tag, where a morphological feature has its own fixed position, the tags for Russian are conditional - the sequence of features depends on the part of speech. This difference, however, is not relevant to us as we leave the morphological tags untranslated, focusing rather on the syntax and the deep syntax transfer.

The nodes are connected between each other with the arcs that are marked with one of 78 syntactic relations (Predicative, Attributive, Adverbial etc.) One of the main "surface" differences from PDT is that the SynTagRus does not regard punctuation marks as nodes, whereas in the PDT analytical (syntactic) level punctuation symbols have even their own syntactic function.

2.3 Data for our experiment

For a parallel treebank we have chosen a part of a Russian novel "Kafedra" ("The Faculty") by I. Grekova, because this novel was also translated into Czech and 480 sentences of it were annotated within the SynTagRus. Those sentences formed the core of our treebank. Probably more sufficient from a point of view of sentence correspondence will be translations of news articles, but they do not exist. We disposed only the printed version of the book which we scanned and aligned the sentences in the text manually.

The main challenge to handle the corpus is its novel translation into Czech. A sentence translated into Czech sometimes bears only a meaning of a source Russian sentence, and it is rather difficult to make the word alignment.

This problem is also multiplied by free word-order of those two Slavic languages. First we supposed that this common syntactic feature will contribute to the similarity of sentences. Afterwards we have found out that while translating the free word order Czech construction, in the Russian sentence the words can be mixed up in another way.

3 Format Transfer for Russian data

SynTagRus is coded in an XML-based format, which we transformed into a PML (Prague Markup Language) format. It would be also rather straightforward to transfer Russian morphological tags into Czech. Morphological systems of the two languages are almost similar, both Czech and Russian have the same cases except for Vocative in Czech, verb tense system is also very close.

On the other hand if we want to be consistent, we should also make a transformation of syntactic properties (Russian syntactic functions) into afuns (Czech analytical functions). Here we face a big problem, because the two annotation schemes have different principles of annotation in this case. There are more than 78 syntactic functions in SynTagRus and only 28 afuns on the analytical layer in PDT, most of which can be mapped into those from SynTagRus (Predicative, Adverbial, Auxiliary relations).

Still, we should not forget about th information on the tectogrammatical layer for Czech, or functors. We argue, that the combination of an analytical function and a functor for Czech can correspond in some cases to a syntactic function from SynTagRus. In other words, the syntactic layer of annotation for Russian is more deep and semanticalized, and it is one layer. Whereas the Czech annotation draws a distinction line between syntax and semantics, leaving syntactic features to the analytical layer and semantics to the tectogrammatical one. This fact and some possible solutions of this problem can be illustrated by an example of a verb argument structure. For instance, in SynTagRus the complement relations are described as syntactic functions “n-compl”, where n is a sequence number of an actant. In the Czech PDT it can correspond to either tectogrammatical functor “Patient” or “Means”. In order to capture such differences we wrote a set of rules, for example they can be schematized as follows:

```
Ru:  1-compl in Accusative case → Cz:  Patient,
Ru:  1-compl in Instrumental case → Cz:  Means.
```

The rules of transfer are now currently under development, and we have found corresponding functors in Czech for all syntactic relations in Russian. More information on the format transfer between the treebanks can be found in [4].

4 Parsing the Czech text

One of the biggest challenges of our work was to annotate the raw Czech data on all the levels - morphological, syntactic and a bit semantic, so that these sentences can be "comparably" aligned to their high-quality manually annotated Russian counterparts. Translated Czech sentences were automatically analyzed using TectoMT framework [10].

The following steps were done:

- tokenization,
- tagging and lemmatization using Morce tagger [12],
- parsing with McDonald's MST parser [5],
- automatic conversion to tectogrammatical trees using mainly rule-based scripts, which are included in TectoMT framework.

Obviously, mistakes in automatically parsed Czech trees occurs. The unlabeled accuracy of the Czech parser is about 85%. We plan to fix them manually in the future.

5 Parallel Treebank Compilation

The parallel treebank is represented on three layers: morphological, analytical and tectogrammatical. The size of the treebank is not very big in comparison with treebanks mentioned in Section 1, and we are currently looking for ways to enlarge the corpus. The statistics of our parallel treebank is summarized in the Table 1.

Table 1: Summary of the Treebank size

	Czech	Russian
sentences	480	480
words	5131	5895

Trees are visualized in TrEd editor[1], which is used for the annotation of the PDT. A screenshot of the annotation on all three layers for both Czech and Russian sentences can be seen in Figure 1. Now we will briefly describe annotation layers of the parallel treebank.

5.1 The Morphological layer and the Word Alignment

The morphological layer shows a sentence in Czech and Russian, where the words go in a linear manner, and they have their morphological properties attached. The whole corpus is automatically aligned on the level of words. For this purposes we ran the GIZA++ tool [9] on parallel texts lemmatized both on the Czech and Russian side. The two resulting one-directional alignments were then symmetrized using intersection symmetrization. For better alignment results we added to our small parallel data the Czech-Russian part of parallel corpus UMC [3]. On the sample of 100 sentences we made a manual evaluation of a word alignment quality,

[1] http://ufal.mff.cuni.cz/ pajas/tred/

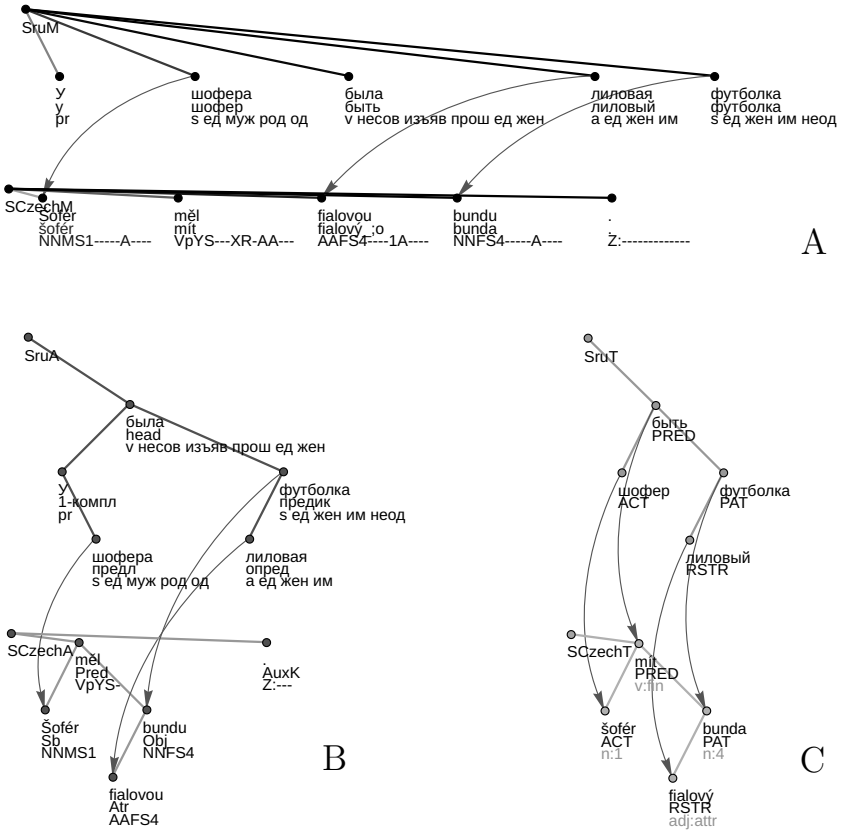

Figure 1: Representation of the sentence "The driver had a lilac coat" for Russian (at the top) and Czech (at the bottom) on morphological layer (A), analytical layer (B) and tectogrammatical layer (C).

its precision reached 85 %. In the future, we plan to improve the word alignment by introducing a good Czech-Russian dictionary.

5.2 The Analytical Layer

The core goal of this project is a task of annotation of the treebank at least on the analytical level, so that syntactic correspondences between the languages can be seen. If not taking into account some surface incorrespondences in Czech and Russian trees caused by different annotation scheme, as, for instance, punctuation marks in Czech scheme which are ignored in SynTagRus, we can compare syntactic constructions in both languages. Figure 2 illustrates a sentence which has more or less similar syntactic structure, and the shapes of two trees are evidently close. In the next section we will show an example of trees with a different syntactic structure.

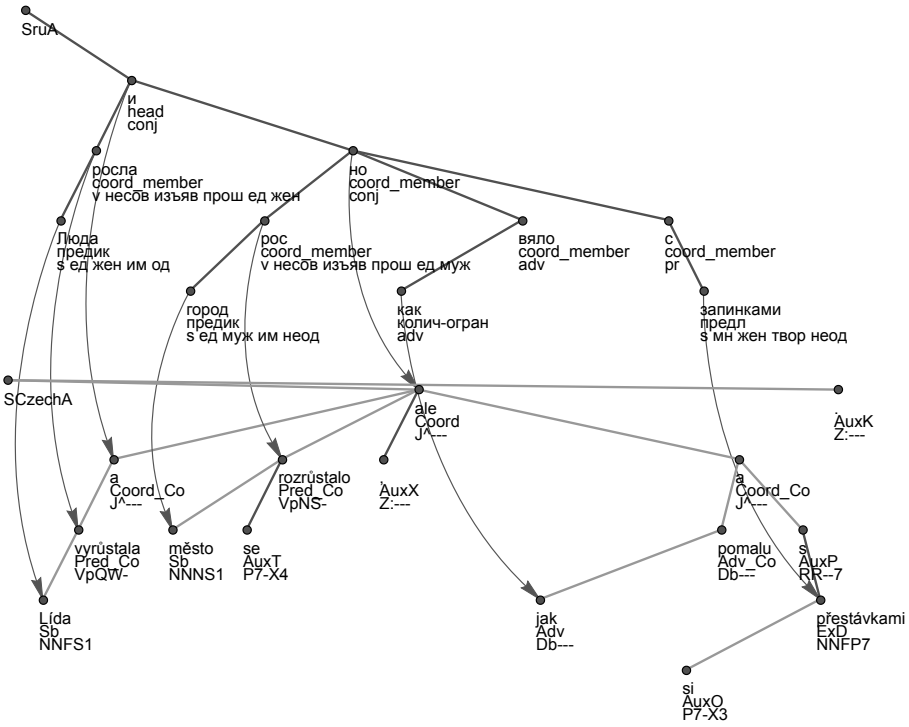

Figure 2: Aligned analytical representation of the sentence (lit.)"Lida was growing up and the town was growing, but somehow slowly, with breaks".

5.3 The Tectogrammatical Layer

The tectogrammatical layer of our parallel treebank is so far annotated only preliminary. It would be a huge work to make correspondences between Czech functors and Russian analytical functions. Still, tectogrammatical trees in two languages will be more similar, than the corresponding analytical trees. One of our tasks for future will be improving the tectogrammatical annotation for this treebank. First insight into the tectogrammatical annotation of Russian is described in [4].

6 Sample Analysis of a Sentence

We have described the preliminary research of how the Czech-Russian treebank can look like. Due to the small size of the parallel treebank it can not be used for the purposes of Statistical Machine Translation, as the PCEDT. However, this annotated data on each of the three layers can bring some insight into the comparative studies that can be useful while designing a Rule-Based or Hybrid Machine Translation system between the languages. As an example of such exploiting for

differences that serve as a basis for MT rules of transfer, we will examine the sentence from the Figure 1.

The *morphological annotation* will provide evidence on whether or not sentences in two languages consist of words with the same or different part of speech, and how similar the morphological properties of those words are. In our example there are four lemmas in Czech and five in Russian (extra one is a preposition).

The *syntactic annotation* can help while inducing basic rules of the syntactic transfer for the Rule-Based MT system. For example, a frequent possessive construction with the verb "to have" in Czech and "to be" in Russian depicted as a tree reflects a difference, which is a candidate for a syntactic rule. To continue, in Czech and Russian sentences the same aligned words have different syntactic functions (“driver” - Subject and a child of the “verb” in Czech, Object and a “child” of the preposition in Russian).

Lastly, two trees on the *tectogrammatical layer* are identical and the corresponding nodes have the same tectogrammatical functors, as this level of annotation stands closer to the “Interlingua”.

7 Conclusion and Future Work

We have shown here the initial phase of building the Czech-Russian Dependency Treebank. On the small sample of the data we made the preliminary correspondence between the two annotation schemes, which will be useful while adding new data to the treebank. One of the possible directions of our research is also making use of automatic annotation tools from the SynTagRus - the tagger and the parser - so that we can annotate a parallel corpus of Czech and Russian languages on syntactic level, not being dependent on the data from the monolingual treebanks. This will enlarge our corpus size at the price of quality, because in addition to the Czech parser mistakes, there will be also mistakes from the Russian parser. The treebank described is not published on-line because of the copyrights. Still, it will be widely exploited for the internal research purposes, namely for constructing rules for the RBMT system between Czech and Russian.

8 Acknowledgments

The research was funded by grants MSM 0021620838, GA201/09/H057, and GAUK 116310/2010. We are also grateful to Leonid Iomdin who provided the Russian data for the project and to the anonymous reviewers for their valuable remarks.

References

[1] Gustafson-Capkova, Sofia, Samuelsson, Yvonne, and Volk, Martin (2007) SMULTRON (version 1.0) - The Stockholm MULtilingual parallel TRee-

bank. http://www.ling.su.se/dali/research/smultron/index.htm. An English-German-Swedish parallel Treebank with sub-sentential alignments.

[2] Čmejrek, Martin, Cuřín, Jan, Havelka, Jiří,Hajič, Jan and Kuboň, Vladislav (2004) Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation *In 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal*.

[3] Klyueva, Natalia and Bojar, Ondřej (2008) UMC 0.1: Czech-Russian-English Multilingual Corpus, *Proceedings of International Conference Corpus Linguistics*, Saint-Petersburg, pp. 188–195.

[4] Mareček, David and Kljueva, Natalia (2009) Converting Russian Treebank SynTagRus into Praguian PDT Style, *Proceedings of Multilingual resources, technologies and evaluation for Central and Eastern European languages*, Borovets, Bulgaria.

[5] McDonald, Ryan, Pereira, Fernando, Ribarov, Kiril and Hajič, Jan (2005) Non-Projective Dependency Parsing using Spanning Tree Algorithms, *Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing (HTL/EMNLP)*, pp. 523–530, Vancouver, BC, Canada.

[6] Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann (1993) Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics 19*, pp. 313–330, [Reprinted in Armstrong, Susan (ed.) (1994) *Using large corpora*, pp. 273–290. Cambridge, MA: MIT Press.]

[7] Mel'čuk, Igor (1988) Dependency Syntax: Theory and Practice, State University of New York Press.

[8] Nivre, Joakim, Boguslavsky, Igor and Iomdin, Leonid (2008) Parsing the SynTagRus Treebank, *Proceedings of COLING08*, pp. 641–648.

[9] Och, Franz Josef and Ney, Hermann (2003) A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics,* **1**, 29, pp. 19–51.

[10] Popel, Martin, Žabokrtský, Zdeněk (2010) TectoMT: Modular NLP Framework, *Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, pp. 293–304.

[11] Sgall, Petr, Hajičová, Eva and Panevová, Jarmila (1986) The Meaning of the Sentence in Its Pragmatic Aspects, Reidel.

[12] Spoustová, Drahomíra, Hajič, Jan, Votrubec, Jan, Krbec, Pavel and Květoň, Pavel (2007) The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech, *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, Praha, pp. 67–74.

Mediating between Incompatible Tagsets*

Alexandr Rosen

Charles University
Faculty of Arts, Prague
E-mail: alexandr.rosen@ff.cuni.cz

Abstract

The issue of incompatible morphosyntactic tagsets in multilingual corpora could be solved by an abstract hierarchy of concepts, mapped to language-specific tagsets. The hierarchy supports the user and tools by resolving categories that do not match the relevant tagset in queries, by providing links between language-specific tagsets, and by displaying responses using a preferred tagset. The hierarchy, built using the methods of Formal Concept Analysis, can also help to refine morphosyntactic annotation in one language by using word-to-word alignments to parallel texts tagged by a different tagset.

1 Introduction

Users of multilingual corpora are often confronted with a variety of language-specific morphosyntactic tagsets. To use tags in a query or to understand its results requires cheat sheets or even lengthy manuals. Without the benefit of intuitive understanding of distinctions and similarities between notationally different or similar tags, multilingual applications drawing on linguistic knowledge and more abstract (syntactic and semantic) annotation schemes built on top of morphosyntactic annotation stumble over an even harder problem.

The ideal solution could be a single consistent standardised annotation scheme in the spirit of *MULTEXT-East* [1]. However, to build a multilingual corpus using such a scheme seems unrealistic, especially when more than a handful of languages are involved.[1] Available taggers are trained on different tagsets, and consistently annotated training data are seldom available even for typologically close languages.

*This work was supported by grant no. MSM0021620823 of the Czech Ministry of Education, Youth and Sports, as a contribution to the parallel corpus project *InterCorp*.

[1]The parallel corpus *InterCorp* currently offers on-line concordances in 23 languages, 14 of them tagged with different morphosyntactic tagsets. The corpus can be queried at korpus.cz/Park after registration at http://ucnk.ff.cuni.cz/english/dohody.php. For more information about the project see http://korpus.cz/intercorp/.

Confronted with texts already tagged in different ways, the user may still believe that tagsets can be translated into a common standard. But a given tag may be too specific or too general to be expressed by a tag from a different tagset. Fig. 1 illustrates the tagset variety using comparable examples of prepositional phrases in 11 languages, tagged by available tools.[2] While some corresponding tags used in the examples are indeed notational equivalents, other tags are not related 1:1. The English tag `IN`, unlike all its prepositional counterparts, is used also for subordinating conjunctions, the German tag `ADJA` covers attributive adjectives (including ordinal numerals) irrespective of degree, while its English counterpart `JJS` is used for superlative adjectives, ignoring the attributive/predicative distinction. The Czech and Polish words *těch* and *tym* are members of the same class, yet the Czech form is tagged as demonstrative pronoun, undistinguished between attributive or substantive use, while the Polish form is tagged on a par with all forms of adjectival declension, including some other types of pronouns and numerals. The partial overlaps in the meaning of corresponding tags are reminiscent of translational mismatches in bilingual dictionaries, including phenomena such as false friends.

en	in `IN`	the `DT`	remotest `JJS`	exurbs `NNS`
de	in `APPR`	den `ART`	abgelegensten `ADJA`	Außenbezirken `NN`
nl	in `600`	dit `370`	schitterende `103`	appartement `000`
fr	dans `PRP`	les `DET:ART`	plus lointaines `ADV ADJ`	banlieues `NOM`
sp	en `PREP`	las `ART`	zonas `NC`	más remotas `ADV ADJ`
it	da `PRE`	queste `PRO:demo`	lingue `NOM`	babeliche `ADJ`
ru	v `Sp-l`	samych `P--pl`	otdaljonnych `Afp-plf`	rajonach `Ncmpln`
cs	v `RR-6`	těch `PDXP6`	nejodlehlejších `AAFP6---3A`	zástavbách `NNFP6---A`
bg	na `R`	tova `Pde-os-n`	prijatelsko `Ansi`	dviženie `Ncnsi`
pl	w `prep:loc:nwok`	tym `adj:sg:loc:m3:pos`	wspaniałym `adj:sg:loc:m3:pos`	apartamencie `subst:sg:loc:m3`
hu	a `ART`	szép `ADJ`	katalán `ADJ`	lányba `NOUN(CAS(ILL))`

Figure 1: Differences in tagging: prepositional phrases

[2] Bulgarian, Dutch, English, French, German, Italian, Russian and Spanish are tagged by *TreeTagger* (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/), Czech by *Morče* (http://ufal.mff.cuni.cz/morce/), Polish by *TaKIPI* and *Morfeusz* (http://nlp.ipipan.waw.pl/TaKIPI/), Hungarian by *HunPOS* (http://code.google.com/p/hunpos/). The tags used here and below are often truncated for brevity.

When the problem of converting between incompatible tags and tagsets concerns only closed-class items (pronouns, function words), it can be solved by using lexeme-specific information corresponding to the source tag (see [6]). In cases involving open word classes we could use an intermediate representation that allows for underspecification at the cost of leaving the target tagset with a potentially imprecise translation of the source tag, as in *Interset* [9]. In the context of many different languages and tagsets, the latter option is more appealing, provided that the language-specific tagsets are correctly linked with the abstract interlingual categories and the representation allows for an arbitrary level of specificity. Both of these features, not inherent to *Interset*, are important for using the representation as the common tagset, and for deriving the most appropriate target tag, which may be too general or too specific, but the extent of the residual part is always known.

Our goal is to delegate the task of dealing with multiple tagsets in a corpus to such an abstract interlingual hierarchy of linguistic categories, where each language-specific tag is mapped onto a node, positioned appropriately with respect to the interpretation of other tags. Because the differences between tagsets often reflect different linguistic perspectives rather than typological distinctions between the relevant languages, a specific word class is seen as an intersection of classification along several dimensions. Following [5] and others, the hierarchy takes three different views of the concept of word class. Thus, the tag for the Czech relative pronoun *který* 'which' is decoded as a category with the properties of lexical pronoun, inflectional adjective and syntactic noun, each with its appropriate morphological characteristics.

Rather than adopting or attempting to design a universal typology of linguistic categories, we prefer to base the hierarchy on distinctions present in our language-specific tagsets and stay open to future extensions. The hierarchy can be built and mismatches between tagsets partially resolved using Formal Concept Analysis [2]. In a parallel corpus with word-to-word alignment and the definition of language-specific domains of the hierarchy, morphosyntactic annotation can be refined by adding information from corresponding tags in other languages, even when the individual tagsets do not make that distinction.

2 Word Classes in 3D

The traditional list of eight word classes is defined by a mix of morphological, syntactic and semantic criteria. For nouns or adjectives the three criteria agree. Nouns refer to entities and decline independently in typical nominal positions; attributive or predicative adjectives represent properties and agree with nouns. On the other hand, numerals and pronouns are defined solely by semantic criteria, while their syntactic and morphological behaviour is rather like that of nouns (cardinals and personal pronouns) or adjectives (ordinals and possessive pronouns). For such cases, the option of a cross-classification along several dimensions seems attractive. Distinctions between the three aspects are borne out also by tagsets. The

Czech tagset has a preference for lexically-based classification [3], the Polish tagset [8] for inflectional word classes, the German tagset distinguishes pronouns by their syntactic function.

A comparison of tags in closely related languages is illustrative. An item tagged as adjective in the Polish tagset (`adj`) can be tagged in the Czech tagset also as an ordinal numeral (`Cr`), possessive (`P8`), demonstrative (`PD`) or relative pronoun (`P4`). A Polish tag for non-inflected words (`qub`) may correspond to a Czech tag for particles (`TT`), non-gradable adverbs (`Db`), reflexive pronouns (`P7`), subordinating (`J,`), or coordinating conjunctions (`J^`).

The 3D space helps to sort out such differences in tagsets. Using the tagset specification, properties of each tag can be identified and related to similar tags in other tagsets. The properties translate into categories in the abstract hierarchy, as in Fig. 2, where the topmost node *wcl* stands for nouns, adjectives and relative pronouns. Its daughters are labelled by a word-class aspect: *lexical* (for 'semantic'), *inflectional* (for 'morphological') and *syntactic*.[3] The other nodes stand for word classes in the three respective dimensions, distinguished in their labels by the initial letter. The seven nodes share only three daughters. Each of the three objects inherits the property of being a word class according to the three criteria.

Each node denotes a set of objects – language-specific tags. The topmost node denotes all tags in all tagsets. Immediate subnodes of a node denote its subsets. A tag denoted by a node must be denoted by at least one of its subnodes. A node can be a subnode of more than one node. In this case, the subnode denotes a subset of the intersection of the sets denoted by its supernodes.

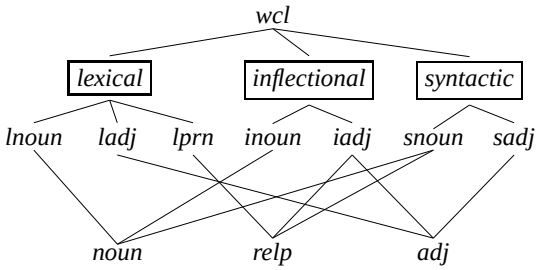

Figure 2: A hierarchy for nouns, adjectives and relative pronouns

Nouns and adjectives are members of their respective classes along all the three dimensions. On the other hand, a Czech *wh-* form *který* 'which' in its use as a relative (rather than interrogative) pronoun (1) is a *syntactic* noun as the subject of the relative clause, a *lexical* pronoun with "dog" as its antecedent, and – due to its adjectival declension – an *inflectional* adjective.

[3] We use *lexical* rather than *semantic* – *lexical* word classes have their properties specified in the lexicon. The boxes around the labels suggest that the sets of objects denoted by the sister nodes are identical.

(1) Psa, který nemá náhubek, do vlaku nepustí.
dog$_{\text{ACC}}$ which$_{\text{NOM}}$ has$_{\text{NEG}}$ muzzle$_{\text{ACC}}$ into train let in$_{\text{NEG,PL,3RD}}$
'An unmuzzled dog won't be allowed on the train.'

The hierarchy in Fig. 3 focuses on Czech numerals and pronouns. ordinals such as *pátý* 'fifth' are treated as *lexical* numeral and adjective – both *inflectional* and *syntactic*. Possessive pronouns differ in being *lexical* pronouns. Personal pronouns are inflectional and syntactic nouns, similarly as cardinal numerals. The interrogative homonym of the relative *který* can be used as a syntactic adjective or noun. The node *intp* inherits from *snom*, representing syntactic nouns *or* adjectives, while *relp* can only be a syntactic noun, due to its ancestor *snoun*.

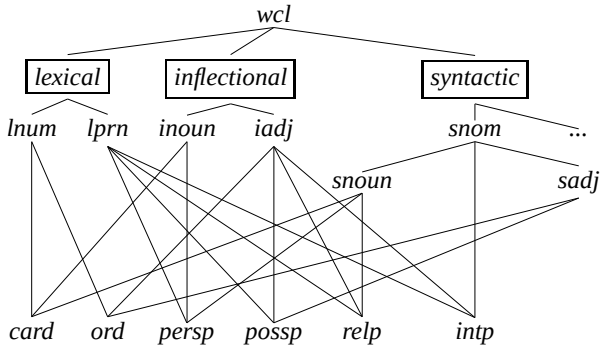

Figure 3: Distinguishing types of numerals and pronouns in a hierarchy

Který in its relative and interrogative use shares a single tag (P4), corresponding to a category ambiguous between relative pronoun and syntactic noun on the one hand and interrogative pronoun and syntactic adjective or noun on the other. The modified hierarchy in Fig. 4 captures this ambiguity. The Czech tag P4 corresponds to a node labelled *lprn* $\wedge$ *iadj* $\wedge$ *snom*.

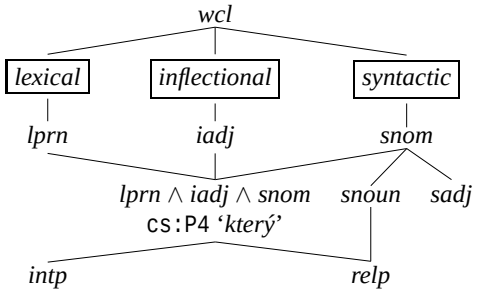

Figure 4: A single node for interrogative and relative pronouns

The concept of three-dimensional word class allows for proper mapping between language-specific tagsets. The tag for adjective in English, German, French,

Italian and Polish covers also ordinal numerals. If all these tags are represented as *syntactic* adjectives, they end up correctly in the same class as Czech, Spanish, Russian or Bulgarian adjectives, ordinal numerals and possessive pronouns. Their *lexical* word class is unknown, although it is not arbitrary. Fig. 5 shows a fragment of the hierarchy with a node representing exactly ordinal numerals and adjectives, labelled $(lord \vee ladj) \wedge iadj \wedge sadj$ and corresponding to the German tag ADJA.

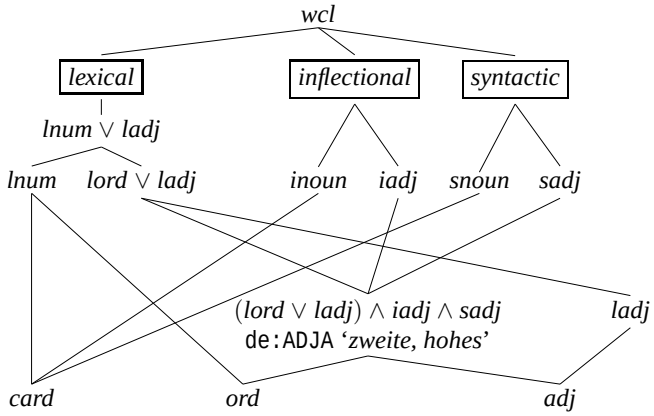

Figure 5: A single node for ordinal numerals and adjectives

The German ordinal number *zweite*, tagged as adjective (similarly as *hohes*), is a subtype of inflectional and syntactic adjective (*iadj* and *sadj*), and also a subtype of a general type covering lexical adjectives and ordinal numerals ($ladj \vee lord$).

Word class of any flavour may be required to co-occur with a set of morphological categories: personal and possessive pronouns with the *lexical* categories of person, number and gender, inflectional adjectives with the *inflectional* categories of gender, number and case. A Czech possessive pronoun such as *jejího* 'her' is *lexically* 3rd person, singular and feminine, while *inflectionally* it is masculine or neuter, singular, genitive or accusative.[4] This is an additional motivation for the three-dimensional approach to word classes.

3 Building and Using the Hierarchy

The hierarchies are equivalent to concept lattices of Formal Concept Analysis (FCA).[5] FCA relates objects according to their attributes with *concept*s, each consisting of a set of objects and attributes as its extension and intension, respectively.

The first step is to identify objects and their attributes in a *formal context*. Table 1 is the formal context for our previous example of adjectives and numerals

[4]Czech personal and possessive pronouns share the same *lexical* categories and are distinguished by their *inflectional* category.

[5]For an overview of linguistic applications of FCA see [7]. [4] is concerned with a lexical interlingua, similar to our hierarchy of linguistic categories.

(Fig. 5). Attributes corresponding to the boxed labels in Fig. 5 are omitted: they would be specified for all objects and would not make the resulting lattice more informative. Next, a set of formal concepts is built. Objects belonging to a concept belong also to its superconcept and the concepts are partially ordered by specificity (roughly: the more attributes, the more specific). Finally, the concept lattice can be drawn (Fig. 6). Its geometry is significantly simpler than the hierarchy constructed intuitively (as in Fig. 5), but the concept ambiguous between adjectives and cardinal numerals is still there. The latter two steps can be done automatically.[6]

	ladj	*lnum*	*iadj*	*inoun*	*sadj*	*snoun*
adj	✓		✓		✓	
ord		✓	✓		✓	
card		✓		✓		✓

Table 1: Formal context for adjectives and ordinal numerals

1	⟨{adj,ord,card},	{}⟩
2	⟨{ord,card},	{*lnum*}⟩
2	⟨{adj,ord},	{*iadj*,*sadj*}⟩
3	⟨{adj},	{*ladj*,*iadj*,*sadj*}⟩
3	⟨{ord},	{*lnum*,*iadj*,*sadj*}⟩
3	⟨{card},	{*lnum*,*inoun*,*snoun*}⟩
4	⟨{},	{*ladj*,*lnum*,*iadj*,*inoun*,*sadj*,*snoun*}⟩

Table 2: Formal concepts derived from Table 1

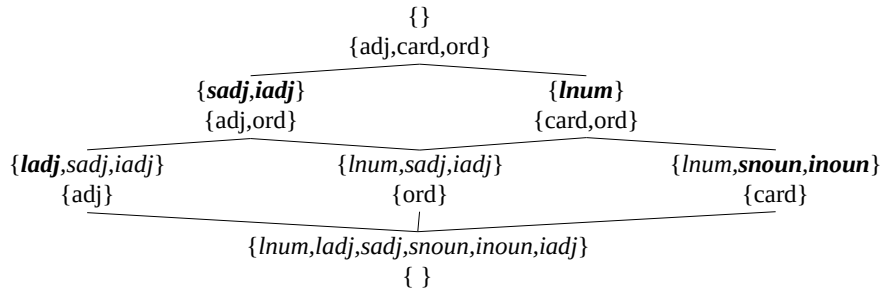

Figure 6: Concept lattice for adjectives and ordinal numerals

Attributes specified for an object in a formal context are interpreted in conjuction. Thus, specifying both *snoun* and *sadj* as attributes of interrogative pronoun (*intp*) would mean that it is syntactic noun and syntactic adjective at the same time. To model disjunction of attributes we have to introduce a more general attribute covering the two options. The formal context for numerals and pronouns is shown below in Table 3 and the corresponding lattice in Fig. 7.

[6] See http://www.fcahome.org.uk/fca.html.

	lnum	*lprn*	*inoun*	*iadj*	*snoun*	*sadj*	*snom*
card	✓		✓		✓		✓
ord	✓			✓		✓	✓
persp		✓	✓		✓		✓
possp		✓		✓		✓	✓
relp		✓		✓	✓		✓
intp		✓		✓			✓

Table 3: Formal context for numerals and pronouns

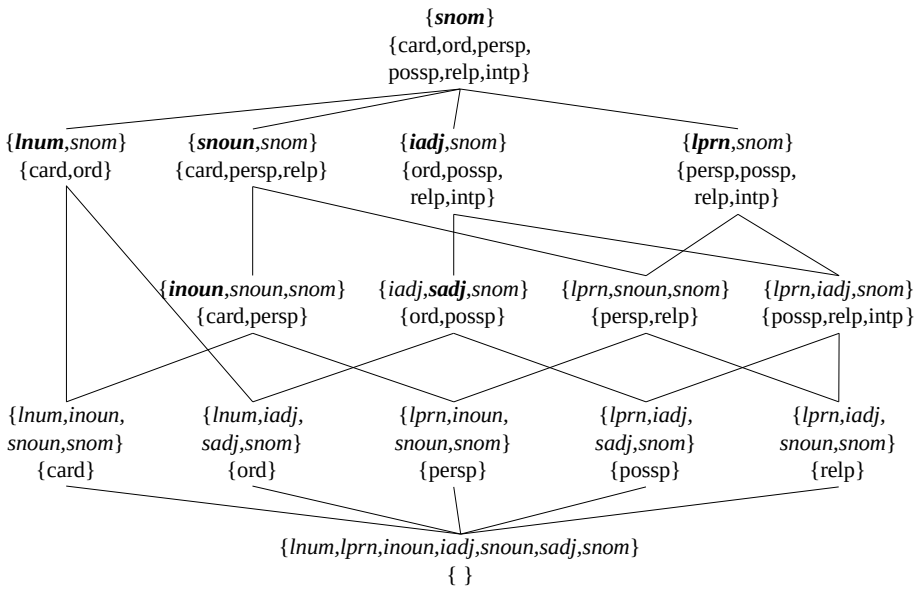

Figure 7: Concept lattice for numerals and pronouns

Lattices can be used for reasoning about attributes, as in the implications *ladj* ⇒ *sadj* or *snoun* ⇒ *lnum*, refering to Fig. 6. Such statements may help the user with language-independent category labels, or to match incompatible language-specific tags. The concept with the extension {ord} corresponds to `Nr`, the Czech tag for ordinal numerals, while the concept with the extension {adj,ord} corresponds to `ADJA`, the German tag covering adjectives and ordinal numerals. Its optimal Czech equivalent would be a Czech tag corresponding to the {adj,ord} concept. In the absence of such a tag, the more specific concepts are traversed and the disjuction of Czech tags corresponding to {adj} and {ord} is the result. Looking up a German equivalent of `Nr` is similar to the scenario when the user asks for "ord" in a German text. It is easy in a Czech text, because the appropriate tag `Nr` is available. For German, there is no tag corresponding to "ord". There are also no concepts more specific than {ord} that would correspond to German tags. The only option is to resort to a more general concept {adj,ord}, with the corresponding German tag.

The extensions of the two concepts can be compared and the user warned that she would have to filter out concordances including categories corresponding to “adj”.

This is a chance for a more data-driven approach to step in. If at least some of the word tokens tagged in the German corpus as ADJA are aligned with their Czech counterparts, the Czech word’s tag may decide whether the German word is a regular adjective or an ordinal numeral. In a multilingual corpus, multiple alignments can be used and a voting scenario applied. Then the hierarchy should decide what kinds of distinctions (i.e. what categories) are relevant for a given language, independently of its tagset.

It seems that incompatible tagsets may actually be useful; there are quite a few cases where projecting morphosyntactic tags in a language pair may bring mutual benefit. In 1.5 million word-to-word alignments extracted from the Czech-English part of *InterCorp*, more than 16.2% of 357 thousand Czech tokens tagged as nouns have their English equivalent tagged as proper noun, which is a category missing on the Czech side. Switching the direction, 85.3% of the total of 95 thousand Czech prepositions have as their English equivalent a token tagged by one of the two highly ambiguous tags: `IN` as preposition/subordinating conjuction or `TO` as preposition/infinitival particle *to*. In 2 million Czech-Polish pairs, 67.2% of 197 thousand Czech tokens tagged as pronouns of different types are likely to have pronominal Polish equivalents, tagged by their *inflectional* class, mostly adjectival or nominal. This opens up the option to project their Czech *lexical* class, although pronouns as a closed class category could be identified as lexemes. The other direction may be more attractive – some Czech pronominal tags are underspecified along the inflectional and syntactic dimensions, which is precisely the information offered by their Polish counterparts. Czech demonstrative and indefinite pronouns (about 31.9% of the total number of Czech pronouns) can thus be identified as attributive or substantive.

4 Conclusion

As a solution to the issue of tagset variety in multilingual corpora we have proposed an abstract interlingual hierarchy of categories, based on a three-way distinction in the system of word classes. In addition to intuitive and underspecified queries and principled mappings between different language-specific tagsets, the hierarchy can be used to refine morphosyntactic annotation in word-aligned parallel corpora by learning from more specifically tagged word tokens in other languages.

If corpus data include only original, language-specific tags, the system can be easily modified and extended without touching the corpus data and the abstract categories can be mapped to tags in any format. Formal Concept Analysis is the answer to concerns about the costs of designing the hierarchy.

The abstract hierarchy is currently built for languages equipped with morphosyntactic annotation and represented in the *InterCorp* project. The work is based on available documentation, annotations actually produced by the taggers,

and previous work, mainly the results of the *Intertag* project. Experiments aiming at the refinement of morphosyntactic annotation by projecting information using word-to-word alignment bring positive results and may be useful even for untagged texts. Although a proper evaluation has not been done yet, it is obvious that incompatible tagsets can actually complement each other and have synergic effects.

References

[1] Tomaž Erjavec. MULTEXT-East Morphosyntactic Specifications: Towards Version 4. In Radovan Garabík, editor, *Metalanguage and Encoding Scheme Design for Digital Lexicography*, pages 59–70, Bratislava, April 2009. Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences.

[2] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis. Mathematical Foundations*. Springer, Berlin/Heidelberg, 1999.

[3] Jan Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague, 2004.

[4] Maarten Janssen. Multilingual Lexical Databases, Lexical Gaps, and SIMuLLDA. *International Journal of Lexicography*, 17(2), 2004.

[5] Miroslav Komárek. Autosemantic Parts of Speech in Czech. In *Travaux du Cercle linguistique de Prague*, volume 3, pages 195–210. 1999.

[6] Natalia Kotsyba, Adam Radziszewski, and Ivan Derzhanski. Integrating the Polish Language into the MULTEXT-East Family: Morphosyntactic Specifications, Converter, Lexicon and Corpus. In *Proceedings of Research Infrastructure for Digital Lexicography: MONDILEX Fifth Open Workshop*, pages 37–55, Ljubjana, Slovenia, 2009.

[7] Uta Priss. Linguistic Applications of Formal Concept Analysis. In Bernhard Ganter, editor, *Formal Concept Analysis. Foundations and Applications*, volume 3626 of *Lecture Notes in Artificial Intelligence*, pages 149–160. Springer, Berlin/Heidelberg, 2005.

[8] Adam Przepiórkowski and Marcin Woliński. A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages, *EACL 2003*, 2003.

[9] Daniel Zeman. Hard Problems of Tagset Conversion. In Alex Fang, Nancy Ide, and Jonathan Webster, editors, *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pages 181–185, Hong Kong, China, 2010.

Annotating the Dutch Parallel Corpus

Hans Paulussen*
ITEC, K.U.Leuven Campus Kortrijk

Lieve Macken†
LT^3, University College of Ghent and Ghent University

Abstract

The Dutch Parallel Corpus (DPC) is a translation corpus containing Dutch, English and French text samples aligned at sentence level. Next to sentence alignment, the corpus has also been grammatically annotated, thus improving exploitation for different domains, including natural language processing, translation research or CALL (computer-assisted language learning). In this paper, we describe the compilation of DPC and the alignment procedures used. This is followed by a description of the annotation task for the three languages, which required different tools and different tag sets. Finally the impact of different grammatical annotations on multilingual corpus exploitation is discussed.

1 Introduction

Over the last decade it has become clear that aligned parallel corpora are indispensable resources for a wide range of multilingual applications. These include different domains, such as machine translation (especially corpus-based MT such as statistical and example-based MT), computer-assisted translation tools, cross-lingual information extraction, multilingual terminology extraction, and computer-assisted language learning.

For some time, high-quality parallel corpora with Dutch as the central language did not exist or were not readily accessible for the research community. This was mainly due to copyright restrictions.

The output of the DPC project is a 10-million-word, high-quality, sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French (Paulussen *et al.* [16], Macken *et al.* [10]). The corpus is a multilingual translation

*email: fistname.lastname@kuleuven-kortrijk.be
†email: fistname.lastname@hogent.be

corpus that not only is aligned at sentence level, but that has also been annotated grammatically and lemmatised for all three languages involved. The combination of aligned sentences and an enriched grammatical annotation layer makes DPC a very useful instrument for multilingual corpus exploitation.

This article starts with a general description of the Dutch Parallel Corpus, pointing out in which way DPC differs from similar parallel corpora. Then the alignment procedures and the annotation procedures used during the compilation of the corpus are described. This is followed by a discussion on the consequences and the impact of the use of different grammatical annotation sets with regard to parallel corpus exploitation.

2 A well-balanced parallel corpus

An important drawback of many parallel corpora is their lack of text balance. For example, the MLCC parallel corpus[1] only covers a selection of the Debates of the European Parliament. Similarly, the EU ACQUIS parallel corpus is solely devoted to European legal texts (Erjavec *et al.* [7]).

One of the main tasks of the DPC project consisted in compiling a well-balanced translation corpus. Therefore, special attention was paid to select a representative sample for each text type. The texts were selected from different domains in compliance with the requirements of the user group. The 10,000,000 word corpus covers translations of the following five text types: literature, journalistic texts, instructive texts, administrative texts and external communication. The texts were selected from different types of text providers including publishing houses, press, government, commercial companies and content brokers (Rura *et al.* [17]).

In order to guarantee the quality of the corpus, a number of validation stages were incorporated during the compilation process, and this for every step of the compilation task: corpus design, text sample selection, cleaning and structuring the data, alignment and annotation of the corpus. For the annotation step, special attention was paid to carefully comply with the annotation protocols proposed by the researchers of the D-Coi project, who compiled a 50-million-word pilot corpus of contemporary Dutch[2].

One of the main tasks of the DPC project consisted in solving all copyright issues (De Clercq and Montero Perez [5]). Thanks to monitoring this delicate task, the corpus is now freely available for the whole research community. The Dutch HLT-agency[3] is in charge of the distribution of the corpus.

[1] URL: `http://www.elda.org/catalogue/en/text/W0023.html`

[2] The D-Coi project also uses the annotation procedures developed for compiling CGN (Corpus Gesproken Nederlands), the Dutch Spoken Corpus.

[3] URL: `http://www.inl.nl/nl/tst-centrale`

3 Sentence alignment

In the DPC project, the alignment was carried out with three different aligners. The basic aligner was the vanilla aligner (Danielsson and Ridings [4]), which is an implementation of the sentence-length-based statistical approach of Gale and Church ([8]). The vanilla aligner has some practical limitations, since it expects texts to have the same number of paragraphs in source and target texts, so that some preprocessing was required.

The second aligner is Melamed's GMA tool (Geometric Mapping and Alignment) ([12]), an implementation of the *Smooth Injective Map Recognizer* (SIMR) algorithm, which is based on word correspondences and sentence length, and relies on finding cognates (tokens with the same meaning and similar spelling) in parallel texts to suggest word correspondences. Optionally translation lexicons can also be used. In the DPC project, we used the NL-Translex translation lexicons ([9]) as an additional source for establishing word correspondences.

The third aligner is the Microsoft Bilingual Aligner developed by Moore ([13]). This aligner uses a three-step hybrid approach to sentence alignment. The aligner uses sentence length and lexical correspondences, both of which are derived automatically from the source and target texts. In a first step, an initial set of high accuracy alignments is established using a sentence-length-based approach. In the second step, this initial set of alignments serves as the basis for training a statistical word alignment model ([2]). Finally, the corpus is realigned, augmenting the initial set of alignments with sentences aligned on the basis of the word alignments. The aligner outputs only 1:1 links and disregards all other alignment types.

During the DPC project we have tested the three aligners, and came to the conclusion that by combining the output of different aligners the amount of manual work necessary to achieve near 100% accuracy can be reduced significantly (Trushkina *et al.* [18]).

4 Linguistic annotation

To improve exploitation facilities of a parallel corpus, an additional linguistic annotation layer was added to the sentence aligned DPC corpus, consisting in grammatical annotation (adding Part-of-Speech tags) and lemmatization[4].

Because of the available tools and the existing PoS tag sets, the job was carried out differently for each of the three languages. Although aiming at complying with annotation standards, such as specified by EAGLES ([6]), we also had to comply with *de facto* standards.

For English, grammatical annotation and lemmatization was performed by the combined memory-based PoS tagger/lemmatizer, which is part of the MBSP tools ([3]). The English memory-based tagger was trained on data from the Wall Street

[4]Note that the sentence alignment task and the annotation task were carried out concurrently. Only a the end of the project, the resulting files were fused into one output format.

Journal corpus in the Penn Treebank ([11]), and uses the Penn Treebank tag set, which contains 45 distinct tags.

For Dutch, the D-Coi tagger was used ([19]). This tagger, which uses the CGN PoS tag set ([20]), is an ensemble tagger that combines the output of different machine learning algorithms. The CGN PoS tag set ([20]) is characterized by a high level of granularity. Apart from the word class, the CGN tag set codes a wide range of morpho-syntactic features as attributes to the word class. In total, 316 distinct full tags are discerned.

```
<seg type="sent" n="seg.p1.s4">
  <seg type="original">
    De fles wordt in geopende toestand met een staalkabel
    in zee neergelaten.
  </seg>
  <s n="p1.s4">
    <w ana="LID(bep,stan,rest)" lemma="de">De</w>
    <w ana="N(soort,ev,basis,zijd,stan)" lemma="fles">fles</w>
    <w ana="WW(pv,tgw,met-t)" lemma="worden">wordt</w>
    <w ana="VZ(init)" lemma="in">in</w>
    <w ana="WW(vd,prenom,met-e)" lemma="openen">geopende</w>
    <w ana="N(soort,ev,basis,zijd,stan)" lemma="toestand">toestand</w>
    <w ana="VZ(init)" lemma="met">met</w>
    <w ana="LID(onbep,stan,agr)" lemma="een">een</w>
    <w ana="N(soort,ev,basis,zijd,stan)"
         lemma="staalkabel">staalkabel</w>
    <w ana="VZ(init)" lemma="in">in</w>
    <w ana="N(soort,ev,basis,zijd,stan)" lemma="zee">zee</w>
    <w ana="WW(vd,vrij,zonder)" lemma="neerlaten">neergelaten</w>
    <w ana="LET()" lemma=".">.</w>
  </s>
</seg>
```

Figure 1: DPC sample annotation Dutch: dpc-bmm-001099-nl

For French, we used a modified version of TreeTagger. This approach was motivated by the fact that the basic PoS tag set of TreeTagger is rather limited for French. Instead of using the basic small PoS tag set, covering only 33 labels[5], we opted for the richer GRACE tag set ([15]). A parameter file based on GRACE was provided by the LIMSI research team, who had created a corpus using the GRACE tag set ([1]). Although this parameter file contained the grammatical information, it did not contain any lemmatized data. In order to solve this problem we opted for a two step annotation cycle. In a first run, the basic parameter file was used and lemmata were added, together with tagging probabilities. The lemmata were then updated and detailed morphosyntactic information was added using a separate tool (FLEMM, [14]). In a second run, the LIMSI parameter file was used, based on the GRACE tag set, covering 312 distinctive tags. Finally, the output of both

[5]The French TreeTagger tagset can be found at URL:
http://www.ims.uni-stuttgart.de/šchmid/french-tagset.html

runs were compared and put together. Only the GRACE tags were retained. The original TreeTagger tags were only used for comparison. The comparison of both runs was also used for spot checking possible errors where the two outputs differ in one way or another.

At the final stage of the DPC project, the output of both main tasks (i.e. sentence alignment and grammatical annotation) were fused together into one XML file, as illustrated in Figure 1, which will be explained in the following section.

5 Discussion

In this section, we discuss the implications of the three PoS tagging formats when used in exploitation of parallel corpora. The tagging codes used are illustrated in example (1), where the token-PoS pairs are given of the Dutch sample shown in Figure 1, together with the token-PoS pairs for the corresponding French and English sentences. A more legible format of the three sample sentences is shown in example (2).

(1) a. De/LID(bep,stan,rest) fles/N(soort,ev,basis,zijd,stan)
wordt/WW(pv,tgw,met-t) in/VZ(init)
geopende/WW(vd,prenom,met-e)
toestand/N(soort,ev,basis,zijd,stan)
met/VZ(init) een/LID(onbep,stan,agr)
staalkabel/N(soort,ev,basis,zijd,stan) in/VZ(init)
zee/N(soort,ev,basis,zijd,stan)
neergelaten/WW(vd,vrij,zonder) ./LET()

b. La/Da-fs-d bouteille/Ncfs est/Vmip3s immergée/Afpfs
ouverte/Vmps-sf à/Sp ses/Ds3fps deux/Ak-fp extrémités/Ncfp
et/Cc fixé/Vmps-sm le/Da-ms-d long/Ncms d'/Sp
un/Da-ms-i câble/Ncms en/Sp acier/Ncms ./F

c. The/DT bottle/NN is/VBZ lowered/VBN into/IN the/DT
sea/NN on/IN a/DT steel/NN cable/NN ,/, open/JJ ./.

(2) a. De fles *wordt* in geopende toestand met een staalkabel in zee *neergelaten*.

b. La bouteille *est immergée* ouverte à ses deux extrémités et fixé le long d'un câble en acier.

c. The bottle *is lowered* into the sea on a steel cable, open.

First we give a brief explanation of the XML-format used for the sample sentences. DPC has been packed in TEI P5 format, in order to have a well-formed and validated format[6]. In the XML-formated samples, a sentence is represented twice.

[6]XML is only considered a wrapping format, in order to distribute the data easily; exploitation of the data can be carried out in whatever format the programmer prefers.

First, the original cleaned sentence is shown in a <seg> element of type original. Then the sentence is shown in an <s> element, including the tokenised format, whereby each word element (<w>) contains two attributes: the PoS tag (ana) and the lemmatized form (lemma). The original sentence can be helpful to reconstruct the original layout of a sentence, when token segmentation is ambiguous. It is also useful for skimming the XML-file quickly, since a flow of horizontally ordered words is easier to read than a list of words displayed in vertical format.

The sample sentences show some general differences, which can be interesting for translation studies. For example, the French version is more verbose than the Dutch and English sentences[7]. In Dutch, typically the auxiliary is placed at the beginning of the sentence whereas the main verb (*neergelaten*) is placed at the end. In French and English, the verb group — shown in italics — remains clustered at the beginning of the sentence. Another example is the word *staalkabel* which is translated as a noun group in English (*steel cable*) and a prepositional construction in French (*câble en acier*). Unfortunately, the underlying categorial labels cannot transparently be matched, which requires some processing. As shown in the three figures, each language uses a different set of PoS tags. In general, this is not such a big problem, but it can be annoying, if you really want to compare grammatical patterns.

When we take a closer look at the PoS tags in the three samples, one can immediately see that categorial and subcategorial information is intertwined in the English tags. In Dutch and French, on the other hand, a more systematic structure is used, which is related to the fact that the two latter languages are morphologically rich when compared to English, but probably also to the fact that English has been the first language for which a tagging scheme has been established.

The Dutch PoS tags have a clear pattern, whereby the tag always starts with the category label indicated in capital letters. The subcategorial features are summed up between brackets. For French, a similar approach is used: the first letter of each PoS tag indicates the grammatical category; all following letters refer to the subcategorial features. Because of the systematic structure of the Dutch and French PoS tags, it is not so difficult to select sentences in both languages based on a categorial filter.

A possible solution to handle the three different tag sets is shown in Table 1. The left column lists the generally accepted 10 basis categories, followed by punctuation labels and miscellaneous labels. The other columns contain the category labels for the three languages. In the case of English, we show the full label pattern, whereas for Dutch and French, only the categorial information is shown. The only exceptions for French are the preposition (Sp) — which is considered an adposition — and the numerals, which are considered a subcategorial feature.

The Dutch and French mapping are closest to the EAGLES proposals, whereas the English tags have a completely different system. The table may be a bit confusing, in the sense that only the English labels show a mixture of categorial and

[7] Note that in French, the bottle seems to be open at both sides: *ouverte à ses deux extrémités*

Basic categories	NL	EN	FR
Noun	N	NNS?, NNPS?	N
Verb	WW	VB[DGNPZ]? MD	V
Adjective	ADJ	JJ[RS]?	A
Adverb	BW	RB[RS]? WRB EX	R
Determiner	LID	DT, WDT	D
Numeral	TW	CD	[NAPD]k Ao
Pronoun	VNW	PRP$?, WP$?	P
Preposition	VZ	IN, TO	Sp
Conjunction	VG	CC, IN	C
Interjection	TSW	UH	I
Punctuation	LET		F
Miscellaneous	SPEC	SYM EX PDT, POS RP, FW LS	X ?

Table 1: Mapping PoS in Dutch, English and French

subcategorial information. Some striking examples for English are the following:

All verb tags, except modal verbs (`MD`) start with `VB`. `WH`-words have a special status, since they are listed as adverb (`WRB`), determiner (`WDT`) and pronoun (`WP$?`). Strictly speaking, these cases are only formal variations of the same category, which only lead to cumbersome selections. Suppose, for example, that you want to check whether a selected sentence starts with a determiner in the three languages, you'll have to specify four different labels (`LID`, `DT` or `WDT` and `D`), which is a bit awkward.

There are two cases which are problematic. The `IN` label is ambiguously considered a *preposition or subordinating conjunction*. The second case is `TO` which is normally used to indicate *to* when linked to an infinitive verb, but which can also be used as an ordinary preposition. Moreover, `TO` as an infinitive indicator has no equivalent in French or Dutch. Depending on the kind of queries you are analyzing, further analysis of the selected material may be necessary.

6 Conclusion

This article presented DPC, a new parallel corpus for Dutch, English and French. DPC is a sentence aligned corpus with an additional linguistic annotation layer, encoding grammatical tags and lemmata. The extra layer makes the corpus more

suitable for fine-grained grammatical selections, at least for monolingual selections. In the case of multilingual selection, things can be rather complicated: the PoS tags differ for the three languages, even when in most cases, these tags are *de facto* standards for monolingual research. Also, the grammatical annotation schemes differ for the three languages.

A partial mapping of the PoS tags is possible, but this is mainly limited to the category level: combination of category and subcategory is not always possible. The labeling system behind the PoS tags for Dutch and French are better structured than the English PoS tags, which is related to the fact that English was the first language to be tagged, but also because English is morphologically poor, in comparison to Dutch or French.

Selection of text samples based on PoS tags are best carried out at category level. Multilingual selections may require mapping of PoS tags. Some tags are ambiguous and my need further analysis.

References

[1] Allauzen, A. and Bonneau-Maynard, H. (2008). Training and Evaluation of POS Taggers on the French MULTITAG Corpus. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco, pp. 28-30.

[2] Brown, P. F., Della Pietra, V. J., Della Pietra, S. A. and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. In *Computational Linguistics* 19(2), 263–311.

[3] Daelemans, W. and van den Bosch, A. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

[4] Danielsson, P. and Ridings, D. (1997). Practical presentation of a "vanilla aligner". In *Proceedings of the TELRI Workshop on Alignment and Exploitation of Texts*, Ljubljana.

[5] De Clercq, O. and Montero Perez, M. (2010). Data Collection and IPR in Multilingual Parallel Corpora: Dutch Parallel Corpus. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010)*, Valletta, Malta.

[6] [EAGLES] Expert Advisory Group on Language Engineering Standards (1996). *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG-TCWG-MAC/R. Version of March, 1996. (URL: `http://www.ilc.cnr.it/EAGLES/home.html`)

[7] Erjavec, T., Ignat, C., Pouliquen, B. and Steinberger, R. (2005), Massive multilingual corpus compilation; Acquis Communautaire and totale, in *Proceedings of the 2nd Language and Technology Conference: Human Language*

Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland.

[8] Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. In *Computational Linguistics* 19(1), 75-102.

[9] Goetschalckx, J., Cucchiarini, C. and Van Hoorde, J. (2001). *Machine Translation for Dutch: the NL-Translex Project*. Brussels/Den Haag, January 2001; 16pp.

[10] Macken, L., Declercq, O., and Paulussen, H. (forthcoming). Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus. In *META*, 56(1).

[11] Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. In *Computational Linguistics*. 19(2): 313-330.

[12] Melamed, D. I. (1997). A Portable Algorithm for Mapping Bitext Correspondence. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL)*, Madrid, Spain, pp. 305-312.

[13] Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the fifth Conference of the Association for Machine Translation in the Americas (AMTA), Machine Translation: from research to real users*, Tiburon, California, pp. 135-244.

[14] Namer, F. (2000). FLEMM : Un analyseur flexionnel du français à base de règles. In *TAL, Traitement automatique des langues*. 41(2): 523-547.

[15] Paroubek, P. (2000). Language resources as by-product of evaluation: the multitag example. In *Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, pp. 151-154.

[16] Paulussen, H., Macken, L., Trushkina, J., Desmet, P. and Vandeweghe, W. (2006). Dutch Parallel Corpus: a multifunctional and multilingual corpus. In *Cahiers de l'Institut de Linguistique de Louvain, CILL*, Louvain-La-Neuve, 32.1-4 (2006), 269-285.

[17] Rura, L., Vandeweghe, W. and Montero Perez, M. (2008). Designing a parallel corpus as a multifunctional translator's aid. In *Proceedings of XVIII FIT World Congress*, August 2008, Shangai, pp. 4-7.

[18] Trushkina, J., Macken, L. and Paulussen, H. (2008). Sentence Alignment in DPC: Maximizing Precision, Minimizing Human Effort. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco.

[19] van den Bosch, A., Schuurman, I. and Vandeghinste, V. (2006). Transferring PoS-tagging and lemmatization tools from spoken to written Dutch corpus development. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-06)*, Genoa, Italy, 2006.

[20] Van Eynde, F., Zavrel, J. and Daelemans, W. (2000) Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, pp. 1427-1434.

Tree Alignment through Semantic Role Annotation Projection

Tom Vanallemeersch

K.U.Leuven, Belgium
Centrum voor Computerlinguïstiek
E-mail: tallem@ccl.kuleuven.be

Abstract

Translation divergences are a challenge for MT and alignment. In this paper, we investigate whether an alignment method based on semantic knowledge improves over approaches for linguistically uninformed word alignment and purely syntax-based tree alignment. We annotate sentences with rolesets from PropBank and NomBank (verbal and nominal predicates and their semantic roles), and link predicates to their auxiliary words (auxiliary, modal and support verbs) using parse trees. We study two language pairs, English-French and English-Dutch. As no extensive semantic resource is available for French and Dutch, the annotation strategy we choose is cross-lingual semantic annotation projection, combined with automatic SRL. A manual evaluation of our system on an English-Dutch sample shows our system is successful at adding links for predicates to the output of a word alignment system (GIZA++) and two tree alignment systems (Lingua-Align and Sub-Tree Aligner). The performance for role linking is significantly lower, due to errors in the English or target parses.

1 Background

Translations tend to diverge from source texts, in different ways and by different causes. Some divergences (also called "translation shifts") are caused by linguistic constraints, others by extralinguistic factors. As Habash et al. [6, p. 85] state, "a translation divergence occurs when the underlying concept or 'gist' of a sentence is distributed over different words for different languages". They mention several divergence types, such as categorial (change of part of speech), conflational (translation of two words by one, e.g. *dar puñaladas* 'give stabs' into *stab*), structural (e.g. addition of preposition to argument of verb) and thematic (switch of subject and object during translation). Tense and aspect are also expressed in divergent ways in languages, involving affixes (*mangeait*), auxiliary verbs (*has eaten*) and periphrastic constructions (*is going to eat*).

Divergences are a challenge for MT and alignment. In the case of MT, the system needs to make the right choices when generating the translation. In the case of alignment, both basic word alignment (linguistically uninformed SMT) and advanced forms of subsentential alignment like parse tree alignment have difficulties aligning divergent structures. Consider the alignment of the following English-Dutch sentence pair in the Europarl corpus (Koehn [9]) in Figure 1. For the sake of clarity, the start and end of the sentences and the child nodes of two non-terminal nodes are not shown. The word-for-word translation of the Dutch sentence is 'a similar objection can be in-brought against'.

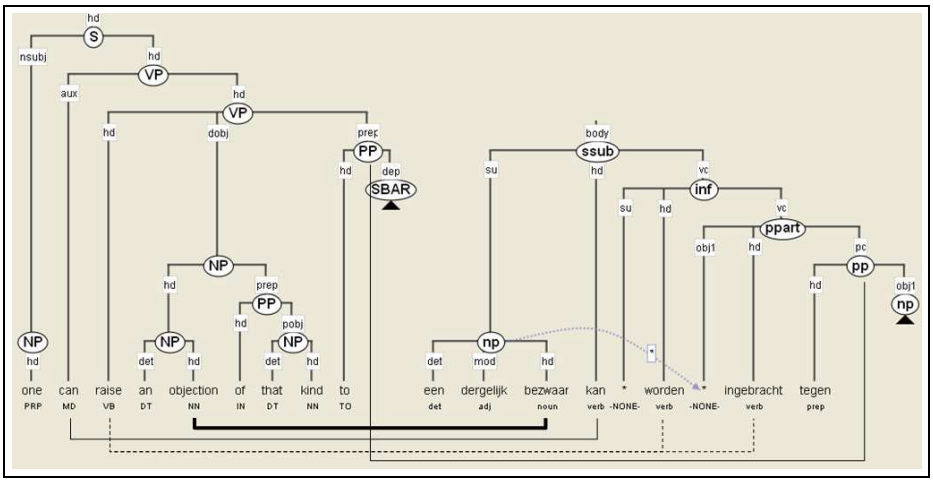

Figure 1: Alignment of divergent structures

The picture shows links created by two variants of the GIZA++ word alignment system (Och and Ney [13]), the highly precise "intersective" and the more extensive "grow-diag-final-and" variant, and by two tree alignment systems, Lingua-Align (Tiedemann and Kotzé [16]), and Sub-Tree Aligner (Zhechev and Way [18]). Only one link is established by all systems (marked in bold at the bottom of the picture). The thin solid links at the bottom of the picture are links that are only procuced by some systems, and no system produces the dashed links (one system aligns one of the words incorrectly). The dashed links involve a Dutch auxiliary of the passive (*worden*) and support verbs of the nominal predicates *objection* and *bezwaar*, which have an argument *to what is merely a report* and *tegen enkel een verslag*. The highly different morphology of the parse trees complicates tree alignment.

In order to tackle translation divergences, semantically oriented approaches have been followed in rule-based MT systems like Eurotra (Allegranza et al. [1]), for coding the argument structure of verbs and for coding tense and aspect in a language-independent way, in order to reduce the transfer step between the two languages to a minimum. In the last decade, automated semantic role labeling (SRL) using one of the available semantic frameworks has become increasingly

important. The idea of semantic roles was pioneered by Fillmore [5], who posited the existence of case relations occurring at deep structure, as opposed to the surface structure of the sentence. Recently, SRL has been applied for multilingual purposes, i.e. to the domain of SMT (Wu [17]) and to parallel text annotation (Padó [14]). In this paper, we propose an approach applying semantic knowledge to the alignment of parse trees.

2 Research Question

Our research question is whether semantic knowledge can improve the alignment of words and constituents. Our assumption is that semantic knowledge is helpful in overcoming syntactic differences between sentences, thus improving over word alignment systems which use linguistically uninformed methods and over methods aligning constituents based on syntactic knowledge only.

The type of semantic knowledge we focus on consists of verbal and nominal predicates and their semantic roles, and the link between predicates and their auxiliary words. The latter are words expressing tense, aspect, modality and passive voice in case of verbal predicates, and support verbs in case of nominal predicates. We only study nominal predicates which are derived from a verb (deverbal nouns). For our purposes, we consider any verb connecting a nominal predicate to one of its arguments as a support verb (see the example *raise - ingebracht* in section 1). The alignment procedure we propose links predicates (and their auxiliary words) and semantic roles between sentences.

The languages we study are English, French and Dutch. This choice was motivated by the fact that English is a resource-rich language and that we want to investigate more than two languages as semantic knowledge is supposed to be applicable beyond a single language pair.

3 Method

In the following subsections, we describe the type of semantic roles we use, our procedure for annotating the predicates and semantic roles and our procedure for determining auxiliary words of a predicate.

3.1 Choice of semantic framework

There are several frameworks for semantic roles, such as FrameNet (Baker et al. [2]), VerbNet (Kipper Schuler [10]), PropBank, (Palmer et al. [15]), and NomBank (Meyers et al. [11]). They are different on many levels, such as coverage, scope of semantic roles, syntactic categories covered, and motivation for their creation. For instance, the motivation for creating PropBank was to annotate predicates and semantic roles in a full corpus, and to train a SRL system on the annotated sentences. Instead of adopting the “traditional” semantic roles (also called theta roles), such

as Agent, Theme and Experiencer (which are used for instance in VerbNet), PropBank marks the roles of verbs as proto-agent (A0) or proto-patient (A1), or as A2, A3 or A4. There are also Argument Modifier roles which apply to any verb and are similar to adjuncts (e.g. AM-TMP for temporal adjunct).

There are links between the different semantic frameworks. The PropBank strategy has been applied to nouns in the NomBank project. When a noun is linked to a verb (deverbal noun), the appropriate roleset of the verb (predicate + roles) is indicated in the NomBank database. In the SemLink project (Loper et al. [7]), PropBank predicates have been linked to VerbNet theta roles, which resulted in a partial mapping between both frameworks. Note that, for alignment purposes, we consider the link of constituents with the same theta role to be stronger than if they only have the same PropBank role.

All semantic resources mentioned above have been initially created for English. Some of them are also available for other languages. For French and Dutch, no extensive resources are available; for Dutch, there exists a limited set of manually annotated PropBank annotations (Monachesi et al. [12]), a rule-based SRL system, and a SRL system trained on manually annotated data. We decided to adopt the PropBank framework for French and Dutch because of the availability of the limited Dutch resource, the framework's aptness for SRL system training, its coverage, and the fact that it covers both verbs and nouns (through NomBank). As creating an extensive PropBank resource for French and Dutch is very time-intensive, we opt for the method of cross-lingual semantic role projection, in combination with the use of a SRL system. This is the topic of the following subsections.

3.2 Projection of predicates and semantic roles

Projection of information from one language to another through alignment links was originally applied to syntactic information (from resource-rich to resource-poor language). Later on, researchers started applying it to fields such as SRL. Padó ([14]) describes an approach for English sentences manually annotated with FrameNet FEEs (frame-evoking elements, which are predicates) and roles, which projects the semantic information to German and French sentences through word alignment links from GIZA++. His primary aim is to study the degree of frame-instance parallellism across languages, i.e. to find out whether the frames used in the source sentences are preserved in the French and German sentences. A number of filters is applied in order to achieve high-precision results and diminish the influence of alignment errors.

Our approach is similar to that of Padó. We apply a SRL system to English sentences, and automated linguistic analysis to the source and target sentences, i.e. parsers that combines constituency and dependency information. We project the predicates and roles to the Dutch and French translation equivalents, using links between words produced by an alignment tool. We then filter out some projections. Our approach differs from that of Padó in the fact that we find links between predicates and their auxiliary words within one sentence and link English predicates

for which no word alignment exists, based on the target parse tree and on auxiliary words of the predicate. The projection procedure is described below. The filters, the detection of auxiliary words and the linking of unaligned English predicates are described subsequently.

The projection procedure is carried out as follows:

- A source predicate is projected to a target token if all of the following conditions are fulfilled: (1) the English predicate is a verb or its roleset has a link to a verb in NomBank, (2) the predicate has at least one non-adjunct role (we ignore Argument Modifier roles for projection), (3) the tokens are linked in the word alignment, and (4) the target token is a non-auxiliary verb or a noun.

- If the source predicate was projected, each of its semantic roles is projected to the smallest constituent from the target parse tree that contains all target tokens linked to tokens from the source constituent. For instance, if the source role A1 is *this particular building* and *this* and *building* are aligned to *dit* and *gebouw*, the role A1 is projected to the constituent *dit gebouw*. We assign a weight to the projection, which is lower than 1 if some of the tokens in the target constituent don't have a link to a token in the source constituent (in the example, the weight is 1). If the weight is too low according to a given threshold, projection of the source role is cancelled.

3.3 Filters on projected information

The first filter removes a predicate (i.e. roleset) if none of its semantic roles was projected.

The second filter checks whether the verb or noun, previously annotated as a predicate through projection, has a direct syntactic connection with the constituents annotated as roles through projection. This filter targets erroneous alignments and strong translation divergences. The filter establishes the shortest path in the target parse tree between the node of the predicate and the node of the role. If none of the nodes in this path is headed by an open-class word (verb, noun, adjective or adverb)[1], the syntactic connection is considered direct. As an exception, we accept one node headed by a verb if the predicate is nominal. An example of the latter can be found in Figure 1: the path between *bezwaar* and *tegen* passes along a modal verb *kan*, an auxiliary verb *worden* and the non-auxiliary verb *ingebracht*. Note that Padó also uses a syntax-based criterion for selecting possible equivalents for a source role, i.e. "argument filtering" (p. 111).

3.4 Linking predicates to auxiliary words

In the three languages under study, there is a limited set of words expressing tense, aspect, modality and passive voice of a predicate. These words, as well as support

[1] With the exception of auxiliary and modal verbs.

verbs of a nominal predicate, are retrieved from the source tree and the target parse tree by checking the sister nodes of the verb and the direct ancestors of the verb. Verbs of modality are only retrieved if they have an infinitival complement. If both the source and target predicate have auxiliary words, we link the auxiliary words to one another. If not, the predicate in one language is aligned both to the predicate in the other language and to the latter's auxiliary word.

3.5 Linking unaligned predicates

In order to overcome weak coverage of the word alignment used for projection, we use two heuristics to link unaligned predicates:

- If a predicate has auxiliary words, and one of those words is linked to a non-auxiliary verb in the target language, we link the predicate to that verb.
- If there is a direct syntactic connection between the projection of an English role and a non-auxiliary verb in the target parse tree, this verb is linked to the predicate of the English role (unless it is already a target predicate in another roleset).

4 Resources

In this section, we describe the resources we apply as input to the method described in the previous section.

We use the Europarl corpus, as it contains translation equivalents in the three languages under study and has been completely parsed and tree-aligned for English, Dutch and French in the framework of the PaCo-MT project (`http://www.ccl.kuleuven.be/Projects/PACO/paco.php`) using Lingua-Align.

As word alignments, we use GIZA++ intersective word alignments.

The SRL system we use is the best-performing system that participated in the CoNLL 2008 task on joint learning of syntactic and semantic dependencies (Johansson and Nugues [8]). It is based on the Penn Treebank and produces syntactic output annotated with PropBank rolesets.

We use parsers which combine dependency and constituency structure:

- English: we convert the syntactic information in the output of the SRL system to Alpino XML format.
- Dutch: we use Alpino (Bouma et al. [3])
- French: we convert the output of the system described by Candito et al. ([4]) to Alpino XML format. This system is trained on a French treebank.

5 Results

We performed an evaluation of our system by running it on a sample of 100 English-Dutch sentence pairs from Europarl, and manually assessing the results. We also compared the results to the output of other alignment systems: the GIZA++ intersective and grow-diag-final-and variants (we used the word alignment produced for the whole Europarl corpus), Lingua-Align (we used the alignment produced for the PaCo-MT project) and Sub-Tree Aligner (which we ran with its standard settings). It should be noted that the Lingua-Align output is based on another English parser than the one we use for semantic projection, i.e. on the Stanford parser. We set the weight for role projection to 0.5.

The SRL system produced 347 rolesets for our sample. After projection and filtering of these rolesets by our system, 150 rolesets remained, corresponding to a total of 444 alignment links (between words or constituents). Table 1 shows the precision for each type of link: links between two predicates, between roles, between two auxiliary words and between a predicate and an auxiliary word.

	number of links	precision
predicates	146	0.95
roles	196	0.83
auxiliary words	35	0.89
predicate-auxiliary	67	0.75
total	444	0.86

Table 1: Alignment precision according to link type

In order to compare our system to the word and tree alignment systems mentioned above, we checked how many links were not present in the other systems and how many links had a different source or target part than in the other systems. Table 2 shows the number of new and different links, and (between brackets) the precision of these links. No figures are given for the role links of GIZA++, as the latter is focused on word alignment.

The precision scores of our system, as well as the comparison with other systems, indicate that our system is highly accurate when aligning predicates, creating links not existing in the other systems, or correcting links of those systems. The system performs significantly less well for roles, especially when we look at the links which are new with respect to the other systems; this is mainly due to errors of the English or target language parser (no efforts were undertaken yet to optimize the weight for role projection). The precision scores for links between predicates in one language (without auxiliary word) and an auxiliary word in the other are also significantly lower than that for predicates. However, these links are helpful in finding predicate links not present in the word alignment (see subsection 3.5).

As far as English-French is concerned, an initial evaluation of our system for that language pair points towards the same conclusions as for English-Dutch.

As an illustration of predicate linking, our system produces the following alignments:

- *you did not call me either - u heeft mij het woord niet verleend* ('you have me the word not provided'): *call* is linked both two *woord* and *verleend* (in all other systems, *call* is only linked to *woord*)
- *the competent services have not included them in the agenda (...) - de (...) diensten hebben die vragen niet op de agenda geplaatst (...)* ('the (...) services have those questions not on the agenda placed'): *included* is linked to *geplaatst* (in all other systems, it remains unlinked)

6 Conclusions and future research

In this paper, we have proposed a method for improving the alignment of words and of syntactic constituents using semantic knowledge. We aim at overcoming syntactic differences between translation-equivalent sentences by determining verbal and nominal predicates and their roles, linking auxiliary words (auxiliary, modal and support verbs) to verbal predicates, and aligning predicates, roles and auxiliary words. The semantic framework we opt for is PropBank, and the annotation strategy is cross-lingual semantic annotation projection, combined with automatic SRL.

The results of our system on a sample of English-Dutch sentences indicate that our system, which is not aiming at a full word or constituent alignment of a sentence pair, is able to improve the output of systems aiming at complete alignment, i.e. a linguistically uninformed word alignment system (GIZA++) and two tree alignment systems based on purely syntactic knowledge (Lingua-Align and Sub-Tree Aligner). Based on the links between predicates, their roles and auxiliary words, and on the information in the source and target parse tree, our system produces highly accurate links between predicates, some of which are not or incorrectly linked in the other systems. On the level of role alignment, the system is

	Lingua	Sub-Tree	GIZA++ int.	GIZA++ gdfa
new pred. links	40 (0.9)	32 (0.91)	25 (0.92)	10 (0.9)
different pred. links	6 (0.83)	12 (1)	2 (0.5)	26 (0.85)
new role links	61 (0.59)	44 (0.52)		
different role links	8 (0.5)	25 (0.8)		
new aux. words	12 (0.67)	5 (0.4)	8 (0.75)	5 (0.4)
different aux. words	1 (1)	2 (1)	1 (0)	2 (1)
new pred.-aux.	52 (0.71)	48 (0.71)	54 (0.74)	28 (0.61)
different pred.-aux.	5 (0.6)	7 (0.71)	3 (0.33)	13 (0.54)

Table 2: Comparison with other alignment approaches (number of links, precision)

less performant, being hampered by errors in the English or target language parser. While the precision scores for links between predicates in one language (without auxiliary word) and an auxiliary word in the other are also significantly lower than the scores for pairs of predicates, these links are also helpful in aligning predicates that are not included in the word alignment.

Our future research involves more extensively evaluating the system output for both language pairs through existing word alignment gold standards, optimizing the role projection threshold and training an SRL system on the annotated target sentences. By running the labeler on the same target sentences, we aim at adding new target predicates and roles to the original ones produced by the cross-lingual annotation projection. New target predicates in a sentence are aligned to source predicates based on the labels of their roles. For the evaluation, we will make use of the existing set of manually annotated PropBank rolesets for Dutch ([12]).

References

[1] Allegranza, Valerio, Bennett, Paul, Durand, Jacques, Van Eynde, Frank, Humphreys, Lee, Schmidt, Paul and Steiner, Erich H. (1991) Linguistics for Machine Translation: The Eurotra Linguistic Specifications. In Copeland, Charles, Durand, Jacques, Krauwer, Steven and Maegaard, Bente (eds.) *The Eurotra Linguistic Specifications*, Office for Official Publications of the Commission of the European Community, Luxembourg, pp. 15–123.

[2] Baker, Collin F., Fillmore, Charles J. and Lowe, John B. (1998) The Berkeley FrameNet Project. In *Proceedings of the COLING-ACL*, Montreal, pp. 86–90.

[3] Bouma, Gosse, van Noord, Gertjan and Malouf Robert (2000) Alpino: Wide Coverage Computational Analysis of Dutch. In *Proceedings of CLIN 2000*, pp. 45–59.

[4] Candito Marie, Crabbé Benoît and Denis, Pascal (2010) Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of LREC-2010*, La Valletta, Malta, pp. 1840–1847.

[5] Fillmore, Charles J. (1968) The Case for Case. In Bach, Emmon W. and Harms, Robert T. (eds.) *Universals in Linguistic Theory*, New York: Holt, Rinehart and Winston, pp. 1–88.

[6] Habash, Nizar and Dorr, Bonnie (2002) Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. In *Proceedings of AMTA-2002*, Tiburon, CA, pp. 84–93.

[7] Loper, Edward, Yi, Szu-ting and Palmer, Martha (2007) Combining Lexical Resources: Mapping Between PropBank and VerbNet. In *Proceedings of IWCS*, Tilburg, the Netherlands, pp. 118–128.

[8] Johansson, Richard and Nugues, Pierre (2008) Dependency-based Semantic Role Labeling of PropBank. In *Proceedings of EMNLP*, Honolulu, Hawaii, pp. 69–78.

[9] Koehn, Philipp (2005) Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, pp. 79–86.

[10] Kipper Schuler, Karin (2005) VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania.

[11] Meyers, Adam, Reeves, Ruth, Macleod, Catherine, Szekely, Rachel, Zielinska, Veronika, Young, Brian and Grishman, Ralph (2004) Annotating Noun Argument Structure for NomBank. In *Proceedings of LREC-04*, Lisbon, Portugal, pp. 803–806.

[12] Monachesi, Paola, Stevens, Gerwert, and Trapman, Jantine (2007) Adding semantic role annotation to a corpus of written Dutch. In *Proceedings of the Linguistic Annotation Workshop, ACL Workshops*, pp. 77–84.

[13] Och, Franz Josef and Ney, Hermann (2003) A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, volume 29, number 1, pp. 19–51.

[14] Padó, Sebastian (2007) Cross-Lingual Annotation Projection Models for Role-Semantic Information. Ph.D. thesis, Saarland University.

[15] Palmer Martha, Gildea, Daniel, and Kingsbury, Paul (2005) The Proposition Bank: An Annotated Corpus of Semantic Roles. In *Computational Linguistics*, 31:1, pp. 71–105.

[16] Tiedemann, Joerg and Kotzé, Gideon (2009) A Discriminative Approach to Tree Alignment. In *Proceedings of the International Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography and Language Learning (in connection with RANLP'09 workshop)*, pp. 33–39.

[17] Wu, Dekai and Fung, Pascale (2009) Can Semantic Role Labeling Improve SMT ? In *Proceedings of EAMT 2009*, Barcelona, pp. 218–225.

[18] Zhechev, Ventsislav and Way, Andy (2008) Automatic generation of parallel treebanks. In *Proceedings of the 22nd International Conference on Computational Linguistics (CoLing)*, pp. 1105–1112.

Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection

Yannick Versley

SFB 833
University of Tübingen
E-mail: versley@sfs.uni-tuebingen.de

Abstract

We present work on tagging German discourse connectives using English training data and a German-English parallel corpus, and report first results towards a more comprehensive approach of doing annotation projection for explicit discourse relations.

Our results show that (i) an approach based on a dictionary of connectives currently has advantages over a simpler approach that uses word alignments without further linguistic information, but also that (ii) bootstrapping a connective dictionary using distribution-based heuristics on aligned bitexts seems to be a feasible and low-effort way of creating such a resource.

Our best method achieves an F-measure of 68.7% for the identification of discourse connectives without any German-language training data, which is a large improvement over a nontrivial baseline.

1 Introduction

Annotation projection is an approach based on using parallel text to transfer linguistic annotations from one language to the other (Bentivogli and Pianta, 2005; Pado and Lapata, 2005); using such techniques, it is possible to bootstrap automatic linguistic annotation for a particular purpose when the respective tools and/or resources are only available in another language – for example, Johansson and Nugues (2006) used such an approach to create a FrameNet parser for Swedish with only a bare minimum of hand-annotation.

In our case, the target consists in explicit discourse relations – discourse relations which are more easy to detect because of the use of so called *discourse connectives*. The category of discourse connectives, despite their common function of linking the contents of two different clauses, is syntactically heterogeneous: It includes coordinating and subordinating sentence conjunctions as the most prototypical examples, but also large and syntactically heterogeneous groups such as multi-word items with conjunction-like behaviour (*as soon as*, *as long as*), and single- or multi-word adverbials that show anaphoric, rather than syntactic, linking

behavior (e.g., *for example*, *in addition*, *on the contrary*). As discourse relations present an abstraction from the concrete (syntactic) means, we expect them to show little variability even in the case of translations that vary in surface word order or syntactic realization, making them an attractive target for annotation projection.

The Penn Discourse Treebank 2.0 (PDTB; Prasad et al., 2008) contains, for the text basis covered by the Wall Street Journal portion of the Penn Treebank, annotation of discourse relations marked by a connective (*Explicit*), those that are not marked by a connective (*AltLex* and *Implicit*), as well as annotations that do not signal a discourse relation (*EntRel* and *NoRel*). About half of the relations in the PDTB are in the *Explicit* category. In contrast to *Implicit* discourse relations, where even a very substantial annotated corpus such as the Penn Discourse Treebank is insufficient for training a reliable automatic classifier, previous research for English has established that finding and classifying explicit discourse relations robustly is well within the reach of the state of the art: Pitler and Nenkova (2009) report an accuracy of 95% for the disambiguation of connective versus non-connective readings of potential discourse connectives, and 94% for classifiying the signaled discourse relation into one of the four top-level categories in the PDTB's taxonomy. Even for the second level of the taxonomy (which is closer to the granularity level found in other discourse-annotated corpora), it is possible to classify instances with about 84% accuracy (Versley, 2011), which is close to the reported inter-annotator agreement for the corpus.

Even for explicit discourse relations, a sufficient amount of annotated data is necessary, as discourse connectives are often ambiguous (between discourse and non-discourse readings, or between different discourse relations), and because the set of discourse connectives is potentially large: The Penn Discourse Treebank contains slightly more than one hundred different discourse connectives; the German Handbuch der Konnektoren (Pasch et al., 2003)), a handbook describing the grammatical properties of German connectives, lists about 300 different connectives. The set of connectives is also syntactically (as well as semantically) heterogeneous, and is not necessarily limited to syntactic constituents. Hence, techniques to reduce the effort for annotating the necessary training examples would be very useful in the creation of discourse-annotated corpora for other languages.

While several medium-to-large discourse corpora exist for English (Carlson et al., 2003; Wolf and Gibson, 2005; Prasad et al., 2008), the availability of resources for other languages including German is much more limited. Among the existing resources for German, the handbook of Pasch et al. (2003) focuses on syntactic properties of different connectives (and would therefore need to be complemented with sense information). Two further resources, the lexicon DiMLex (Stede and Umbach, 1998) and a small RST-annotated corpus (Stede, 2004) have been described in the literature but are not publically available.

Using projection from automatically tagged instances on a parallel corpus, we tackle the problem of bootstrapping the annotation of discourse connectives by investigating (i) the variability in the translation of these items, and (ii) possible approaches to create an automatic tagger for German discourse connectives based on the annotated data.

In order to tag German text using the English training data, several intermediate steps are necessary: Firstly, the original training data has to be used to create automatic annotation for the English side of the parallel corpus (section 2); Second, the annotation on the English side of the parallel corpus has to be projected across the alignment to form training data for the German side of the corpus (section 3). Finally, the projected German data can be used to learn a classifier and annotate a gold-standard sample of German newspaper text (section 4).

2 Tagging of English Connectives

To tag connectives in English text, we use classifiers that are trained on data from the Penn Discourse Treebank 2.0 (PDTB; Prasad et al., 2008). We used an approach that allows fully automatic identification and disambiguation of discourse connectives that is loosely based on the work of Pitler and Nenkova (2009), with modifications that make it more useful for our task: Firstly, our approach creates tags that correspond to the finer second level of the PDTB's taxonomy of discourse relations. Secondly, Pitler et al make use of information that can be found in the hand-annotated treebank, but not in automatic parses (traces and semantic function labels such as '-PRP'), whereas our approach is able to reach similar accuracy (i.e., better by a fraction of a percent) using information that can be derived from automatic parses.

Using the 15 366 *Explicit* relations from sections 2-22 for training data, our tagger is able to distinguish between discourse and non-discourse usages of potential connectives with 92% precision and 98% recall (in cross-validation on the WSJ text); disambiguation of discourse instances between the four coarse relation types in the Penn Discourse Treebank (*Comparison*, *Expansion*, *Contingency* and *Temporal*) is possible with 94-95% accuracy, whereas distinguishing between the sixteen second-level relations (e.g., distinguishing between *Concession* and *Contrast*, *Cause* and *Condition*) is possible with about 84% accuracy. The third and finest level on the PDTB taxonomy can be disambiguated with 79% accuracy. These accuracy results mirror the decrease in annotator agreement reported by Prasad et al. (2008), which leads us to believe that they correspond to a greater difficulty in the disambiguation task (rather than widespread lack of features).

For the experiments, we automatically tagged the English side of the EuroParl corpus using the Berkeley parser[1] for syntactic preprocessing. On a sample of data from EuroParl corpus, we see that the use of automatic parsing and out-of-domain data leads to a slight decrease in performance, with 83% precision and 97% recall.

2.1 Syntax and Tense Features for English Connectives

Two of the features used in our discourse tagger are reimplementations of ones used by Pitler and Nenkova (2009): One is the string of the connective itself, with mod-

[1] http://code.google.com/p/berkeleyparser/

ifiers – such as "*two minutes*" in "*two minutes before the train departed*" omitted in order to ensure the generalizeability of the connective instances.

The second group of features comprises syntactic features, namely the labels of self, parent, left sibling and right sibling nodes (counting from the lowest node that covers all of the words annotated as connective span and that is not the only child of its parent), as well as additional features signaling the presence of a VP node or of a trace as a child of the right sibling.

A third group of features is based on arguments, which we can identify reliably for a restricted subset of the discourse connectives (subordinating and coordinating conjunctions, w-adverbials (*[S . . . [SBAR [WHADVP when] he sleeps]]*). For fronted (preposition- or adverb-headed) adverbials, we can reliably identify one of their arguments, which is the parent clause, sentence whereas the other argument is linked anaphorically and is not identifiably as easily.

Based on the identified arguments, we extract the following indicators:

- the part-of-speech of the first non-modal verb in the sentence (descending from the argument clause node into further VP and S nodes to cover both nesting of VPs and coordinated sentences)
- the presence (and word form) of modals and negation in the clause
- a tuple of *(have-form, be-form, head-POS, modal present)* as proposed by Miltsakaki et al. (2005).

(In the result tables, the part-of-speech/presence of modals pair of features will be called *pos/md*, whereas the tuple describing auxiliaries, the POS of the lexical head, and the presence of modals will be simply called *verb*).

Verb tense and modals are relatively shallow correlates of more interesting properties such as facticity or veridicality (i.e., whether the speaker asserts the propositional content of that clause to be true), but they are easy to extract in a robust manner and useful as a first approximation to a more comprehensive approach such as those of Palmer et al. (2007) to classifying situation entities.

3 Mapping Discourse Annotation

In order to create annotations on the German side, we have to project the automatically annotated data from the English side using available sentence and word alignments to create training data for a German classifier. In a simple first step, we can create a projected version of the English annotation by simply considering every token that has a word alignment link to an English connective;[2] in the case of discontinuous connectives (English *if/then* or *either/or*) or of discontinuous word alignments, the resulting connectives on the German side can be discontinuous.

[2]The word alignments themselves are postprocessed from the statistical alignments that are output by GIZA++, using the *intersection* or *grow-diag-final* heuristic. The experiments where the heuristic used is not stated use *grow-diag-final* since the intersective alignments are often too sparse.

Directly using the projected data on the German side can be problematic not just because of the noise from the statistical word alignment, but also due to a mismatch between the noisy alignments and the syntax-based mechanism used in the English connective finder: The approach chosen for English is dictionary-based in that a list of potential connectives is used to identify candidates by looking for occurrences of the particular word sequences, and subsequently using binary classification (based on syntactic and tense features) to filter out non-connective occurrences.

Dealing with the problem of noise in the projected annotation is possible in two different ways: One would be to deal with the problem by using a shallower sequence tagging approach in the connective classification that does not use a pre-established list of connectives; the other would be to use a pre-established list of connective candidates (i.e., word sequences that can have a connective function, but may be ambiguous between discourse and non-discourse readings), either from an external source such as the HdK, or induced by refining the connective candidates that can be extracted from the word-based projection.

3.1 A Simple CRF Baseline

As a baseline, we consider the most straightforward way to do annotation projection and learn a proposal mechanism: We project the discourse connective annotation on the English side using the word alignments (tagging all German words that are aligned to an English discourse connective), and use the tagging created through this method to train a sequence classifier.

Besides the words themselves, the sequence classifier uses features signaling the start and end of clauses (from automatic parses), and the types of those clauses.

This alignment approach creates relatively noisy annotation, as not all discourse connectives are present in the German translation. Witness the following example:

(1) *{Das}* *ist* *ganz* *im* *Sinne* *der* *Position,* *die* *wir* *als* *Parlament*
{That} is wholly in terms of the position, which we as Parliament
immer *vertreten* *haben.*
always advocated have.
[Indeed], it is quite in keeping with the positions this House has always adopted.

The English *Indeed*, which would signal that the sentence is an explanation of the previous sentence, is not present in the German translation. As a result, an arbitrary part of the sentence is aligned to the discourse connective and receives the connective span. The result is still useful, though, if the classifier that is learned somehow averages out the noise that occurs in training. The alternative to living with this noise, though, is to look for ways to improve the precision, as in the two following approaches.

3.2 A Dictionary-based Approach

An alternative to plain sequence tagging would be an approach more similar to the dictionary-based approach for English, where potential connectives are extracted using a (monolingual) list of such items and these candidates are filteres using a binary classifier (into actual discourse connectives, and word sequences that look like a discourse connective, but actually are not).

In the initial step, we use the list of connectives contained in the German *Handbuch der Konnektoren* to identify potential discourse connectives in the German EuroParl text; if multiple overlapping occurrences are found (e.g., *als*/*when* vs. *als ob*/*as if*), the longest match is kept.

To find out whether a given occurrence should be treated as a positive or as a negative example, we compare its span with all sets of words projected from potential connectives on the English side and use an overlap metric (Dice) to determine which potential connective string on the English side corresponds best. If the best match potential connective on the English side is tagged as a discourse connective, the German span is used as a positive example; if it is not aligned to a potential connective on the English side or the aligned string is tagged as not being a discourse connective, the German span is used as a negative example.

The subsequent binary classifier uses a language-independent version of the syntactic features that are used in the English-only classifier: the connective string, and features describing the lowest common node in the parse tree (label of self, parent, and left and right siblings). For the syntactic preprocessing of German trees, we use the parser of Versley and Rehbein (2009), with a grammar learned from the TüBa-D/Z treebank (Telljohann et al., 2009).

3.3 Inducing a Connective List

While the HdK provides us with a list of connectives, it is an interesting and potentially useful question whether we can induce such a list from the aligned data. As all word alignments have been created automatically, and translators occasionally omit or add discourse connectives in sentences, however, we have to correct or filter the word sequences that can be extracted from the alignments.

For each candidate string, we determine the following three statistics:

- the *total* number of occurrences
- the number of occurrences that overlap with a projected discourse connective (i.e., where at least one word of the candidate string is aligned to at least one word from the English discourse connective)
- for each aligned occurrence, a Dice-based overlap measure between the tagged English discourse connective and the projection of the candidate string (where 0 means no overlap and 1 means that they cover exactly the same words).

Using a dataset composed of the HdK list and a random sample of other candidate strings (both limited to those that had at least 15 aligned occurrences), we found out that the most effective method to discriminate between connectives and spurious candidates was to require a minimum average overlap of about 66-70% (over all aligned occurrences).

To build the list, we took all proposed strings that had at least 15 aligned occurrences, where the average overlap was at least 70% and where the product of (i) the average overlap and (ii) the ratio between aligned and unaligned occurrences was not smaller than $\frac{1}{25}$. These occurrences were then ordered by average overlap (considering better-overlappig proposed strings first) and discarding any proposed string where a subsequence had a higher average overlap.[3]

The resulting list contains 293 items, of which some are not in the HdK list, either as new connectives that fit the HdK criteria, or as items that would need to be manually corrected or filtered.[4]

4 Evaluation and Discussion

To evaluate performance on the German side, we annotated a text sample comprising slightly more than 5000 tokens of text from the TüBa-D/Z corpus, with two annotators independently performing the annotation and merging the differences, yielding 136 connective instances.

The annotated gold standard reflects the criteria set forth in the German *Handbuch der Konnektoren* for grammatical properties of a connective x:

- x cannot be inflected.
- x does not assign case to elements in its syntactic environment.
- x realizes a binary relation.
- The arguments of x are propositional.
- The arguments of x are clauses.

As can be seen in table 1, simply tagging every string from the HdK's list (*all HdK* as a discourse connective results in very good recall[5] but also poor precision.

[3] Keeping shorter proposed strings in the list does not change the end result much, since the tagging process will prefer longer matches over shorter ones.

[4] When ranked by average overlap, the first 62 candidate strings have a high proportion of connectives that are also part of the HdK (69%), some new items (18%, e.g. *anders ausgedrückt* in addition to *anders gesagt* as equivalent to *in other words*)), some are truncated (e.g., *facto* instead of *de facto*), or contain additional tokens such as commas or complementizers (10%), and some which do not fit the criteria for a discourse connective at all (3%). At the bottom of the list, the overlap with HdK items is substantially lower (29%), while the proportion of incomplete/longer items (35%) as well as incorrect items (19%) are much higher. The proportion of correct items not covered by the HdK stays about the same (17%).

[5] Note that the recall is not 100% since we found phrases that match the HdK's criteria for connectives, but are not part of the handbook's list.

	Prec	Recl	$F_{\beta=1}$
all HdK	27.0	**94.9**	42.1
simple CRF, giza-refined	74.0	41.9	**53.5**
simple CRF, giza-intersect	**83.9**	38.2	52.5
HdK+classifier	62.3	76.5	**68.7**
induced+classifier	58.3	56.6	57.5
HdK+CRF	74.7	43.4	54.9
induced+CRF	70.2	43.4	53.6

Figure 1: Evaluation results: Tagging German text (newspaper sample)

For the CRF approach, we used Léon Bottou's Stochastic Gradient Descent CRF learner[6] using default settings (50 training epochs, C=1.0). The CRF baseline yields a much better precision (both for the intersected alignments for *giza-intersect* and using the *grow-diag-final* heuristic for *giza-refined*) but relatively poor recall around 40%.

Using word alignments and the HdK word list we can also derive a binary classifier for occurrences of potential connectives. Such an approach gives a precision that is significantly better than the pure dictionary-based approach, with a comparably smaller loss in recall (76.5% against 94.9%, which is however still considerably better than the 41.9% reached by the CRF-based approach).

To establish whether the improvement in the dictionary-based approach is to be seen in the cleaner training data, or in the more expressive features that are used in the syntax-based classification, we performed additional experiments to reflect the utility of these modifications in isolation. One experiment uses the syntax-based approach with an induced lexicon instead of the HdK one (*induced+classifier*), which results in a substantial loss in comparison to the manually annotated list, but still visibly better results than for the CRF approach.

In contrast, using the CRF approach with training data derived in a different way – using a dictionary in addition to word alignments, and removing anything that cannot be mapped to an entry in the list – shows only very little improvement over the CRF-based method where raw projections were used.

4.1 Summary

In this paper, we presented an approach to transfer a tagger for English discourse connectives by annotation projection using a freely accessible list of connectives as the only German resource. Compared to the supervised approach of Dipper and Stede (2006), who reach 78% F-measure on positive instances for a selected sample of nine German connectives, our annotation projection approach fares reasonably

[6] http://leon.bottou.org/projects/sgd

well.

Ongoing work will concentrate on three main issues: One issue is to complement the annotation projection of discourse connectives with mechanisms to find their sense (i.e., the discourse relation they signal), as well as their arguments. While the mechanisms for argument finding as well as for sense disambiguation that are used for English should in principle also work with other languages, German annotation for these features is not available yet.

The second main issue consists in the word alignments we have used (heuristically refined results from GIZA++), which are admittedly geared towards use in machine translation rather than being optimized for linguistic quality. Since discourse connectives most often consist of function words (rather than content words, which are easier for unsupervised alignment), the alignment of discourse connectives is especially quality-sensitive. Quite possibly, using a more elaborate approach, such as the reordering approach of (Collins et al., 2005), or more comprehensive procedures, such as the direct alignment of parse nodes (Zhechev and Way, 2008; Tiedemann and Kotzé, 2009), can further improve the quality reached by the approach.

A third broad issue is the creation of more expressive features on the German side, including tense/mood-based features, which have been shown to be beneficial for English tagging of discourse connectives.

Acknowledgements The research reported in this paper was financed by the Deutsche Forschungsgemeinschaft (DFG) as part of Collaborative Research Centre (SFB) 833 "Constitution of Meaning". The author would like to thank Emily Jamison and the three anonymous reviewers for helpful comments on earlier versions of the paper and to Anna Gastel for performing part of the connective annotation.

References

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. *Natural Language Engineering*, 11(3):247–261.

Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current Directions in Discourse and Dialogue*. Kluwer.

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause reordering for statistical machine translation. In *ACL 2005*.

Dipper, S. and Stede, M. (2006). Disambiguating potential connectives. In *Proceedings of the Konvens-2006 Workshop on the Lexicon-Discourse Interface*.

Johansson, R. and Nugues, P. (2006). A FrameNet-based semantic role labeler for Swedish. In *Proceedings of Coling/ACL 2006*, pages 436–443, Sydney, Australia.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A., and Webber, B. (2005). Experiments on sense annotations and sense disambiguation of discourse connectives. In *TLT 2005*.

Pado, S. and Lapata, M. (2005). Cross-lingual projection of role-semantic information. In *Proceedings of HLT/EMNLP 2005*.

Palmer, A., Ponvert, E., Baldridge, J., and Smith, C. (2007). A sequencing model for situation entity classification. In *ACL 2007*.

Pasch, R., Braußе, U., Breindl, E., and Waßner, U. H. (2003). *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin / New York.

Pitler, E. and Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *ACL 2009 short papers*.

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 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Stede, M. (2004). The Potsdam Commentary Corpus. In *ACL'04 Workshop on Discourse Annotation*.

Stede, M. and Umbach, C. (1998). DiMLex: A lexicon of discourse markers for text generation and understanding. In *Coling 1998*.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., and Beck, K. (2009). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

Tiedemann, J. and Kotzé, G. (2009). Building a large machine-aligned parallel treebank. In *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories (TLT'08)*.

Versley, Y. (2011). Towards finer-grained tagging of discourse relations. In *Beyond Semantics: Corpus-based investigations of pragmatic and discourse phenomena (Workshop at the annual meeting of the DGfS)*. to appear.

Versley, Y. and Rehbein, I. (2009). Scalable discriminative parsing for German. In *Proc. IWPT 2009*.

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

Zhechev, V. and Way, A. (2008). Automatic generation of parallel treebanks. In *Coling 2008*.