Converting an English-Swedish Parallel Treebank to Universal Dependencies

Lars Ahrenberg
Linköping University
Department of Computer and Information Science
lars.ahrenberg@liu.se

Abstract

The paper reports experiences of automatically converting the dependency analysis of the LinES English-Swedish parallel treebank to universal dependencies (UD). The most tangible result is a version of the treebank that actually employs the relations and parts-of-speech categories required by UD, and no other. It is also more complete in that punctuation marks have received dependencies, which is not the case in the original version. We discuss our method in the light of problems that arise from the desire to keep the syntactic analyses of a parallel treebank internally consistent, while available monolingual UD treebanks for English and Swedish diverge somewhat in their use of UD annotations. Finally, we compare the output from the conversion program with the existing UD treebanks.

1 Introduction

Universal Dependency Annotation (UD) is an initiative taken to increase returns for investments in multilingual language technology (McDonald et al., 2013). The idea is that a common set of dependency relations, and a common set of definitions and guidelines for their application, will better support the development of a common cross-lingual infrastructure for the building of language technology tools such as parsers and translation systems.

UD actually comprises more than just dependency relations. To be compatible and possible to merge in a common collection, the resources for a language should use the same principles of tokenization, and common inventories of part-of-speech tags and morphological features. UD advocates a conservative approach to tokenization, which treats punctuation marks and some clitics as separate tokens, but treats all spaces as token separators. Thus, multiword expressions are not recognized as such until the dependency layer.

For parts-of-speech a tag set comprising 17 different tags only is recommended with a basis in the twelve categories proposed by (Petrov et al., 2012). For an overview, see Table 2 in section 3.

LinES (Ahrenberg, 2007) is a parallel treebank currently comprising seven sub-corpora (see Table 1). Future plans for LinES include a substantial increase in the amount of data included. This would also entail that new contents would not, as a rule, be manually reviewed. Harmonizing its markup with that of other treebanks would make it possible to develop more accurate taggers and parsers for it, and thus increase its usefulness as a resource. Conversely, the monolingual treebanks can be used to augment other treebanks for English or Swedish as training data for parsers and taggers.

Source	Segments	EN tkns	SE tkns
Access help	595	10451	8898
Auster	788	13512	13337
Bellow	604	10310	9964
Conrad	622	13063	12092
Europarl	594	9334	8715
Gordimer	756	15181	15778
Rowlings	605	10299	10635
Total	4564	82150	79419

Table 1: LinES corpora before conversion.

The primary aim of this work is the creation of a UD-compatible version of LinES, LinES-UD. As far as possible this should happen through automatic conversion. The hypothesis is that LinES markup is sufficient to support automatic conversion to universal dependencies for both languages by the same process.

The paper is organised as follows. The next section reports related work. Section 3 presents the primary differences between the design of the LinES treebank and the UD framework. In section 4 we describe our approach to develop the conversion program, and in section 5 we present and discuss the results. Section 6, finally, states the conclusions.

2 Related work

Universal Dependencies is a project involving several research groups around the world with a common interest in treebank development, multilingual parsing and cross-lingual learning (Universal dependencies, 2015). The annotation scheme for dependency relations has its roots in universal Stanford dependencies (de Marneffe and Manning, 2008; de Marneffe et al., 2014) and the project also embraces a slightly extended version of the Google universal tag set for parts-of-speech (Petrov et al., 2012). At the time of writing treebanks using UD are available for download from the LINDAT/CLARIN Repository Home for 18 different languages (Agić et al., 2015).

The first release of UD treebanks included six languages. Two of these, the ones for English and Swedish, were created by automatic conversion (McDonald et al., 2013). The English treebank used the Stanford parser (v1.6.8) on the WSJ section of the Penn treebank for this purpose. The Swedish Talbanken treebank was converted by a set of deterministic rules, and the outcome is claimed to have a high precision "due to the fine-grained label set used in the Swedish Treebank" (p. 93). The treebanks are divided into three sections for the purposes of parser development, a training part, a development part, and a test part. We refer to them in the sequel as the English UD Treebank (EUD) and the Swedish UD Treebank (SUD), respectively, using suffixes 1.0 and 1.1 to differentiate the versions. They have been used extensively in the current project for comparisons. In the most recent release (1.1) some corrections have been made to both treebanks. As far as the syntactic annotation is concerned, the corrections affect less than 1% of the tokens in EUD, and about 4% of the tokens in SUD. Most of the development work on LinES-UD was made with the previous versions as targets, but the comparisons reported in section 5 refers to the versions 1.1.

Several other UD treebanks have been developed as a result of automatic conversion, e.g. for Italian (Bosco et al., 2013), Russian (Lipenkova and Souček, 2014), and Finnish (Pyysalo et al., 2015). The process used here for LinES is quite similar to these works with the special twist that here two parallel treebanks are converted simultaneously. Thus, the approach is rule-based, although the rules are not available in an external rule format, but implemented as conditions and actions in a Perl script. Also, unlike these works no new language-specific UD-scheme is developed as part of this work, as such schemes exist for English and Swedish already.

3 Differences in design

The original LinES design has several differences from the UD treebanks. The differences pertaining to parts of speech are fairly small, while differences in sentence segmentation, tokenization and dependency analysis are larger.

We first observe that parallel treebanks are often created for different purposes than mono-lingual treebanks. UD treebanks have parser development as a primary goal, while the most important purpose of the LinES treebank is as a resource for studying the strategies of human translators and for testing properties that are sometimes claimed to be typical for translated texts. One way to describe the relation between a translation and its source text is by trying to quantify the amount of structural changes, or shifts, that have been performed. Such a task is obviously helped by using the same annotation scheme for both languages and the demands on consistency in application of the categories are high. A measure of structural change should reflect real differences; if they instead are introduced by alternative schemes of tokenization or by the use of different categories or definitions, the value of the measure is reduced.

Some of the differences in the available English and Swedish UD treebanks will be detailed in section 4. Here we only note that they pose problems for a developer of parallel English-Swedish treebanks. As just said, in a parallel treebank we would like to see parallel constructions be annotated in the same way for both languages, but if they are not annotated this way in the (usually much larger) available monolingual treebanks, the increase in parsing consistency that we expect from training the parser on a union of UD-

treebanks, will not be as large as it could be.

3.1 Sentence segmentation

The largest syntactic unit in LinES is a translation unit. This means that it should correspond under translation to a similar unit in the other language. When the translator has chosen to translate one English sentence by two Swedish sentences, or two English sentences by one Swedish sentence, LinES treats the two sentences as a single sentential unit sharing a single root token. From the monolingual perspective there are two sentences, each with its own root, but from the bilingual perspective there is a single unit and a single root. The two sentences can be analysed as either being coordinated or one being subordinated to the other; in the first case one token that would be taken as the root from the monolingual perspective is assigned a conjoining relation to the other root, while in the second case the dependency would be adverbial. An example of a 1-2 alignment is given below, where the root verb of the second Swedish sentence, *skedde* corresponding to 'was' is seen as conjoined to the root verb of the first sentence, *varit*, corresponding to 'been'.

EN: *As Olivia said, it ought to have been a sad-feeling place but it wasn't; there was instead a renewal: ...*
SE: *Det borde, som Olivia brukade säga, ha varit ett dystert ställe men var det inte. Tvärtom skedde en förnyelse: ...*[1]

We note also that some punctuation marks such as the colon or the semi-colon are sometimes treated as sentence delimiters and sometimes not, even in monolingual treebanks. For example, in the English UD corpus the colon sometimes occur in mid-sentence and at other times at the end of sentences.

3.2 Tokenization

LinES treats a number of fixed multiword expressions from closed parts-of-speech categories as single tokens. English examples are mostly complex prepositions and adverbs such as *because of, after all, instead of, in spite of* while Swedish also has multiword determiners such as *den här* (this) and *den där* (that). Although they are not very numerous, some 10% of all sentences would contain a multiword token. As the tokenization principles for UD favours a strict adherence to spaces as separators, instead signalling multiword expressions in the dependency annotation, the conversion to UD must retokenize the data.

The treatment of clitics in LinES are largely the same as in UD with one exception, the English s-genitive. This is treated as a separate token in the English UD treebank, but in LinES it is taken as a morpheme, both for English and Swedish. While arguments can be given to treat the s-genitive as a phrasal clitic also in Swedish, it is usually not done, because it is harder to detect in Swedish than in English.

In LinES hyphens are regarded as token-internal characters. This is not the case in English UD, where many hyphens are treated as separate tokens.

3.3 Parts of speech

The inventory of parts-of-speech in LinES comprises 23 categories. Many of them correspond more or less directly to those used in UD, but there are a few differences. See Table 2 for an alignment of LinES part-of-speech labels to UD labels. The most problematic difference is that LinES makes a differentiation between verbs and participles, whereas UD distributes participles on the categories VERB, ADJ and NOUN. For the current conversion program we have chosen a simple mapping that does not consider all possible variation to determine what it should be converted to. When used as an attribute it is interpreted as an adjective, but in all other cases it is categorized as a verb.

Auxiliaries, including forms of the verbs *be* and its Swedish counterpart *vara*, are another issue. In LinES there is no distinct part-of-speech for auxiliaries; instead the distinction between auxiliaries and ordinary verbs is made on the basis of whether they participate in a verbal chain or not.

A third issue is the distinction between determiners and pronouns. In LinES a word is classified as a determiner only when it introduces a noun phrase. In UD, however, the distinction is not made in the same way. Rather than identifying the individual words that need re-categorization, we have kept the distinctions as in LinES.

[1] The source text is 'A Guest of Honor' by Nadine Gordimer, translation into Swedish by Magnus K:son Lindberg.

POS	EUD	SUD	LinES
ADJ	Yes	Yes	A, PCP
ADP	Yes	Yes	PREP, POSP
ADV	Yes	Yes	ADV
AUX	Yes	No	V
CONJ	Yes	Yes	CC, CCI
DET	Yes	Yes	DET, A, PRON
INTJ	Yes	Yes	IJ
NOUN	Yes	Yes	N, PCP
NUM	Yes	Yes	NUM, ORD
PART	Yes	Yes	ADV, INFM
PRON	Yes	Yes	PRON, POSS
PROPN	Yes	Yes	PN
PUNCT	Yes	Yes	FE, FI, FP
SCONJ	Yes	Yes	CS
SYM	Yes	No	SYM
VERB	Yes	Yes	V, PCP
X	Yes	Yes	No

Table 2: UD Part-of-speech tags, their application in EUD and SUD and their counterparts in LinES.

3.4 Dependency relations

The set of dependency relations in UD currently includes 40 relations; the exact number seem to change every now and then. For example, (de Marneffe et al., 2014) lists 42.

LinES uses 24 dependency relations which are largely based on those used in FDG or Functional Dependency Grammar (Tapanainen and Järvinen, 1997), but with some additions required by LinES corpora and some amendments. As in UD the dependencies largely favour content words to be governors, but not to the same extent. In LinES prepositions are heads, not just case markers, and in constructions with a copula + predicative, the copula is taken to be the head rather than the head of the predicative. For conversion to UD, then, these relations must be reversed, not just relabelled, which in turn may cause structural changes of other kinds. A reversal implies that dependents of the previous governor must be reanalyzed and a decision be made whether they should keep with the previous governor or become dependents of the new governor. For instance, in LinES annotation a copula can have both a subject dependent and adverbial dependents, while in UD all of these dependencies should be transferred to the predicative head.

One reversal may also affect the outcome of another reversal as when the object of the preposition is a clause with a copula, as in *Kim wanted to talk about how stupid I was*. Here, the mapping introduces a direct dependency between two tokens that previously only were indirectly related (see Figure 1).

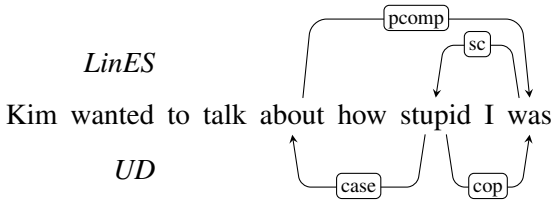

Figure 1: A reversal of governance affecting another. LinES relations above the sentence and UD relations below.

UD largely employs different dependency relations for different parts of speech, whereas LinES prefers to treat dependency relations as orthogonal to parts-of-speech. For example, in LinES there is a single subject dependency which applies to nominals as well as clauses or verb phrases, and a single object dependency applying to nominal as well as clausal dependents. In UD, on the other hand, nominal dependents are consistently assigned different relations than clausal dependendants, whether they are in a subject, complement, or modifier position. Similarly, modifiers are analysed differently as nominal (nmod), adjectival (amod), adverbial (advmod) or numerical (nummod).

LinES shares with UD the assumption that the first conjunct of a coordinated constructions should be the head. In UD all other conjuncts are then taken to be dependents of this first one, whereas in LinES they are (as in FDG) chained so that the next one in the chain is taken to be a dependent of the previous one rather than the first one. Chains of auxiliaries are treated similarly; the first one in a chain of auxiliaries becomes a dependent of the next one, rather than on the main verb, i.e., the head of the last auxiliary, as is the case in UD. Also in agreement with FDG, the subject is a dependent of the first (finite) auxiliary in LinES whereas it is a dependent of the main verb in UD.

LinES provides no dependency information for punctuation marks. The part-of-speech information is however more specific than the single category PUNCT used by UD.

LinES dependency graphs are strictly projective. There are special relations signalling that the dependency should actually not be with the head

assigned, but with some other token, usually a (direct or indirect) dependent of the assigned head. There is one relation for fronted elements, one for postposed elements and one for noun-phrase-internal relations. The situation in UD is not quite clear; on the one hand there seem to be a desire to avoid non-projective relations as the relation 'dislocated' seems to relate a fronted or postposed element to the head of the clause. The relation 'remnant' as used by (de Marneffe et al., 2014) to handle ellipsis, is clearly non-projective, though.

The structural differences provide more or less of a challenge to conversion. Luckily, not all differences involve changes to the dependency structure. Many relations are apparently the same except possibly for the label. In other cases, and unlike the situation with subjects and objects, LinES actually has more specific relations than UD. For example, in LinES a difference is made between prepositions that introduce an adjunct and those introducing a complement (i.e., oblique objects), which is not made in UD. In the same vein, LinES separates adverbial modifiers of verbs from those modifying adjectives, and adjectival modifiers appearing before and after a head noun. For these cases conversion basically means relabelling.

4 Method

The descriptions and examples provided on (Universal dependencies, 2015) have been used to learn the intended meaning and use of the relations. Both English and Swedish pages have been consulted. Although this information is indicative rather than complete, and leaves a lot to the reader's interpretation, we decided that it would be sufficient for a first version of a conversion program. In addition we used the English and Swedish UD treebanks, EUD and SUD, made available by the UD consortium as references for comparing the output of our conversion program.

As we noted above it is important that the two halves of a parallel treebank are internally consistent in their annotation. Now, while both EUD and SUD are UD-conformant, there are differences in how they have applied UD. Thus, it was not possible to make LinES-UD internally consistent and at the same time make its English half consistent with EUD and its Swedish half consistent with SUD. In each case where there is a difference, we had to make a decision which one to follow.

Some of the differences between EUD and SUD are listed in Table 3. First we note that EUD employs a few more dependency labels than SUD. The following labels used in EUD are not found in SUD1.1: *conj:preconj, det:predet, goeswith, list, nmod:npmod, nmod:tmod, remnant*, and *reparandum*. On the other hand, SUD has one label, *nmod:agent*, not used in EUD. We decided to use the dependency labels found in SUD, including *nmod:agent*, as LinES has a special relation for agents in passive clauses.

Aspect	**EUD**	**SUD**
No. of pos tags	17	15
No. of dep. labels	45	38
Hyphens can be tokens	Yes	No
Negation as PART	Yes	No
's as own token	Yes	No
subj/dobj determiners	Yes	No

Table 3: Major differences relating to application of UD in the English and Swedish UD treebanks.

As for parts-of-speech we used the 17 categories found in EUD, although symbols (SYM) and unassigned (X) are quite rare in the corpus. For each language a small set of auxiliary verbs are assigned the category AUX. We also followed EUD in classifying the negation as PART(icle) and possessives as PRON(ouns) for both languages. However, in other aspects LinES UD is closer to SUD: hyphens are not separate tokens and determiners can not be subjects or objects. In the case of genitive -s, we decided to follow EUD for English, making it a separate token, but SUD for Swedish where it is taken to be a morpheme. This actually contradicts our desire to be internally consistent, but was made nevertheless.

4.1 Development phases

The conversion program has been developed iteratively in three phases. The goal of the first phase was to create UD-conformant annotations for all dependencies appearing in the LinES data. A first version was developed for one of the seven sub-corpora, and when the result appeared to be fairly complete, it was tested on the other six. The output was checked for remaining LinES-annotations. When this happened, the cause was quite often an annotation error in the LinES input file, which could be corrected. At other times defaults were introduced.

In the second phase the full LinES treebank was

used. To check for progress frequency statistics were collected on part-of-speech tags, dependency labels and their associations. Agreement with the EUD and SUD was checked by counting triplets of dependency label, dependent part-of-speech and head part-of-speech. A surprising observation was the large number of labels assigned to any given part-of-speech pair. As an example, see Table 4, where frequencies for dependency relations relating an adjective to a head noun are given. At least 18 dependency relations have instances for this pair in either EUD1.0 or LinES-UD. Where frequencies are low one can suspect that we are actually dealing with errors, either in the source data or in the conversion process.

Dependency	**EUD1.0 Frequency**	**LinES-UD Frequency**
amod	3198	3334
acl:relcl	31	0
conj	22	37
nmod	18	34
acl	9	108
case	8	1
appos	5	10
nsubj	5	2
compound	3	0
nmod:npmod	3	0
parataxis	3	0
advmod	2	6
det	1	214
advcl	1	2
nmod:poss	1	0
nummod	1	0
root	0	1
compound:prt	0	1

Table 4: Distribution of dependencies involving an ADJ(ective) as dependent and a NOUN as head in the English UD Treebank and the English half of Lines-UD after conversion. A subset of EUD1.0, selected so as to produce the same total number of dependencies as LinES-UD, was compared with the output of the conversion program.

When differences were striking, the reason was investigated by looking at a sample of instances, and a decision was made whether to change the program in some respect, or leaving it in that stage, usually for the reason that internal consistency between the English and Swedish parts of LinES were judged to be more important than agreement with the UD treebanks. The most striking difference in Table 4 concerns the relation *det*, where LinES-UD have 214 instances and EUD 1. This is explained by the fact that a number of common words that can be termed adjectival pronouns, such as *another, many, other, same, such* are treated differently in the two treebanks, either at the part-of-speech classification (e.g. *another* is DET in EUD, ADJ in LinES) or at the dependency classification: adjectives are regularly analysed as *amod* in EUD, while they can have a *det*-dependency in LinES.

Another difference is the number of 'acl:relcl'-relations for the pair ADJ - NOUN which is non-existing in the output from the conversion program. This turned out to be a miss in the program: relative clauses without relative pronouns or complementizers were not recognized.

When frequency statistics seemed to be fairly reasonable a manual review (by the author) was performed on 50 English and 50 Swedish segments. The results, all around 90%, are shown in Table 5. Apart from a rough quantitative measure of accuracy the review revealed several types of recurring errors in the output, necessitating a third phase of improving the conversion program.

4.2 The conversion program

The program takes three arguments: source and target files in XML-format and their associated alignment file. It returns monolingual files in conllu-format and a new alignment file.

Structure is as a rule handled before labels. The first structural change concerns tokenization. All multiword tokens in LinES have been split into their parts and the word alignment files have been updated accordingly. At the same time, the new tokens are assigned a new part-of-speech (from a specially designed word list) and an appropriate dependency relation, usually 'mwe' except for some multiword proper names, where 'name' is used. The new tokenization requires a renumbering of the tokens of the treebank, and consequently, a renumbering of the links. The total increase in number of tokens is about 0.9%.

Before the changes in the dependency structure are tackled, the part-of-speech mapping is performed. This is motivated by the fact that tagging usually precedes parsing and that it involves no loss of information, as all information pertaining to parts-of-speech or morphosyntactic features

Corpus	Tokens	UAS	LAS
LinES-UD SE	891	0.93	0.90
LinES-UD EN	959	0.91	0.88

Table 5: Accuracy (unlabelled and labelled) of the generated annotations for a small random sample of output from the conversion program.

Type of change	EN	SE
Relabelling	57891	54781
Reversal	9113	9511
Swap	5718	6726
Combination	61	84
Addition	10026	8662
Total	82809	79764

Table 6: Structural mappings and their frequencies in the conversions to LinES-UD. A change of governor is a Reversal if the new governor was previously a direct dependent, a Swap if it was not, and a Combination if it involves two reversals, as in Figure 1. Additions apply only to punctuation marks.

in LinES-corpora can still be accessed by the program. Most of the mapping is just relabelling, either one-to-one or many-to-one, but, as noted above, the category PCP (for participle) is mapped onto three UD tags using contextual information and the verbs are divided on the two categories AUX and VERB depending on whether they are part of a verbal chain or not.

The final step deals with the dependency tree. A new tree is generated from the existing one on the basis of rules that refer to dependency labels, local structure and properties of the two tokens related in the dependency. The more complex structural changes, i.e., reversals and swaps (head changes), are handled first. The given sentence is read three times, first to look for structural changes, then to handle relabellings, and finally to handle punctuation marks.

(Bosco et al., 2013) makes a distinction between 1:1 and 1:n dependency mappings; both of these types are handled as relabellings. The difference is that 1:n mappings, such as the splitting up the LinES object relation on the various corresponding UD dependencies (dobj, iobj, ccomp, xcomp), require inspection of the available morphosyntactic information and local properties of the tree to be performed correctly. In the final pass punctuation marks are assigned the relation 'punct' and a head. The UD recommendations have been followed as far as possible, but it is generally quite problematic to identify a proper head, especially for many of the internal punctuation marks that some authors of novels like to employ.

5 Results and evaluation

The conversion program has been applied to the full corpus and as a result a UD-version of the parallel treebank now exists. In fact, several versions have been generated, as the program is still being worked upon. Here we report on stable properties of the output.

The output has been checked for completeness and for the occurrence of dependency relations not belonging to UD. Although a few tokens, usually less than ten for each language, do not receive any dependency relation or a non-UD label, we can claim that the conversion program is successful in producing a parallel UD treebank. Such errors can be detected and fixed in a manual review.

Frequencies of structural mappings of different types are summarized in Table 6. The number of structural changes (reversals or swaps) is quite high, around 20% for both languages, a bit less for English and a bit higher for Swedish.

While the output is formally in agreement with UD relations and part-of-speech categories, there is no guarantee that they have been applied in agreement with their intended definitions. To check for this frequency statistics have been computed for parts-of-speech and dependency labels, and for dependency triplets.

Table 7 shows total number of instances for the most common dependencies for English and Swedish. We have omitted some, such as *list, goeswith*, and *compound*, that are used only for one language or have a low frequency for one language. For most relations the numbers are quite similar, but there are also exceptions. As the four underlying corpora are different, and we don't have a gold standard for either of them, we cannot determine with any certainty whether the differences are due to text properties, language-specific interpretations of the UD labels, or conversion errors.

More detail can be had by looking at frequencies for dependency triplets. Space is not sufficient to discuss all variation in this data, but we will look at a few pertinent cases. First, we can observe (as

Dependency	EUD1.1	EN LinES-UD	SUD1.1	SE LinES-UD
All	82809	82809	79764	79764
punct	10028	10025	8663	8662
case	7638	8157	8448	8284
nmod	6965	7537	7853	7824
det	6282	**8028**	5680	5145
nsubj	5864	**7215**	6234	6992
dobj	3762	3797	3535	4230
amod	3750	3620	3715	3503
mark	3063	2707	2571	**3631**
advmod	2923	**4692**	5165	5969
conj	2633	3276	3439	3603
aux	2627	2492	1996	1934
cc	2372	2529	2831	2981
cop	1456	1250	1294	1246
advcl	1352	1335	1478	1015
nmod:poss	1279	1535	1424	1562
ccomp	**1126**	**549**	436	560
xcomp	1104	1183	876	1204
nummod	1122	296	1172	225
appos	754	564	424	572
acl:relcl	**708**	**253**	1095	853
acl	707	**1598**	571	**966**
auxpass	650	642	39	167
nsubjpass	561	**70**	1121	**354**
mwe	207	382	**1562**	343

Table 7: Absolute frequencies for the most common dependency relations in each treebank. For both EUD and SUD subsets have been used that are of the same size in terms of number of tokens as the LinES treebank. Bold face is used for relations where differences are noteworthy.

in Figure 4) that the association between dependency labels and pairs of parts-of-speech is n-to-m with sometimes very high values on n and m. For instance, looking at all four treebanks there are no less than 93 pairs of part-of-speech with at least one instance of *nmod*. Similarly, there are 62 pairs with at least one instance of *nsubj*. Of course, often only a few pairs contribute to the vast majority of the instances, but there is almost always a long tail of other pairs.

Some differences can be explained with reference to the texts which are taken from different genres. EUD has newspaper (Wall Street Journal) prose, SUD 'professional prose', while LinES has a great share of literary prose. To illustrate, both EUD and SUD have more than three times as many numerals as the LinES corpus, which largely explains the frequency differences relating to *nummod*. Conversely, LinES_SE has ten times as many occurrences of the pronoun *han*, 'he' than SUD.

The *det*-relation is more frequent in LinES-UD_EN than in EUD1.1 for the reasons explained above, namely that it is used for many common words categorized as ADJ, where EUD uses *amod*. Thus, EUD has more instances of *amod*-relations in spite of having a lower relative frequency of adjectives.

LinES_EN has more *nsubj* instances than EUD. This is largely explained by the frequencies of third person singular pronouns as subjects, especially the pronouns *he* and *she* which are used to refer to the characters of the narrative. Together they account for more than 1000 instances of the difference. And to this can be added the pronouns tagged as PRON in LinES but as DET in EUD.

On the Swedish side, SUD has many more instances of NOUN as subject, while the Swedish LinES-UD again has more pronouns. 23.8% of all tokens in SUD are nouns, while the corresponding figure for Swedish LinES-UD is 17.4%. Con-

versely, SUD has only 6.2% pronouns, whereas Swedish LinES-UD has 11.1%.

The higher frequency of *advmod* in English LinES is partly explained by the higher relative frequency of adverbs, 5.5% as compared to 4.1%. In a corpus of 82000 tokens this is a difference of 1200 instances. The number of adverbs in the Swedish translations is even greater, 7.4%.

The difference in frequencies for *ccomp* in the English treebanks could also be explained by the differences in genres. However, while some verbs that take clausal complements, such as *announce* don't occur in LinES, there are no large differences in frequencies for common verbs taking clausal complements such as *say*, *think*, or *know*. Browsing the LinES file for occurrences of these words, no errors are detected, so the tentative conclusion is that they are used differently.

The conversion program identifies fewer relative clauses than it should, judging from the differences in frequency for the relations *acl* and *acl:relcl*. In particular it misses some that are not introduced by a relative pronoun or subjunction.

The very low figures for *nsubjpass* is partly due to the rules creating this dependency, which are too restrictive, for example missing instances where an auxiliary appears between the subject and the passive form. Another contributing factor is the Swedish word *som*, 'that', 'who', which introduces relative clauses. In SUD it is categorized as a PRON(oun) and assigned a core dependency, whereas in LinES it is categorized as a subjunction carrying the *mark*-dependency. Other words that are analyzed as *mark* much more often in Swedish LinES than in SUD1.1 are *när*, 'when', *då*, 'when, as' and *medan*, 'while'.

SUD1.1 has many more instances of the *mwe*-relation than the other treebanks. While EUD and LinES-UD_EN agree on mwe:s, SUD1.1 employs *mwe* for many word sequences that LinES regards as compositional, such as *när det gäller*, 'as regards', *mer än*, 'more than', *i samband med*, 'in connection with'.

While the most common dependency triplets such as <amod, ADJ, NOUN> and <nsubj, NOUN, VERB> appear in the same numbers, there are thus other triplets occurring in one treebank that don't occur at all in the other treebank of the same language. This indicates (i) that a parser trained on one of them might not perform very well on the sentences of the other, and (ii) that merging the treebanks may not be so helpful either. To test these hypotheses we trained Malt parsers on the two Swedish treebanks and tested various models. The LinES data was randomly divided into distinct sets for training, development and test and parsing models were then developed on the training data for both treebanks as well as for the merged treebank. As both Swedish treebanks are small with many tokens occurring in only one of them, the nouns, proper names, verbs and adjectives were de-lexified into combinations of part-of-speech tags and (LinES) morphological tags. The best results, obtained with the standard settings and finegrained de-lexification are shown in Table 8. No combo model from the merged treebank was able to improve performance on both test sets.

Model	**Test data**	**UAS**	**LAS**
LinES	LinES	0.751	0.701
Combo	LinES	0.739	0.687
SUD1.0	SUD1.0	0.738	0.697
Combo	SUD1.0	0.739	0.696

Table 8: Parsing results.

6 Conclusions

We have shown that the information in the LinES parallel treebank is sufficient to produce a treebank by automatic means, which, with a minimum of manual effort, is formally compliant with the UD inventory of dependency labels and part-of-speech categories, and its principles for tokenization. The program generates English and Swedish data, as well as the new alignment, in one go.

The current version is relatively stable, but there is still room for improvements. Even so, a manual review process will increase the quality of the annotation substantially. The conversion programme will facilitate the review process, however, as we can see from the comparisons with the EUD and SUD treebanks, where the problems seem to reside.

We have also shown that EUD and SUD, while UD-compatible, do not treat all phenomena in the same way. Thus, it is likely that future UD treebanks, whether developments of EUD and SUD, or created from other sources, will be more consistent with one another. In such a future scenario, LinES-UD is likely to follow suit and, rather than having to manually review the data once more,

tweaking an automatic conversion program to the new developments will be more efficient.

We have pointed out that a parallell treebank developed for the study of human translation must be internally consistent to a maximal degree. Presently, this can only be achieved to the expense of deviating in many aspects from the available UD treebanks, some of which have been detailed in section 4. A possibility, of course is to maintain two versions of the data. As part of the parallel treebank, the two halves are maximally consistent with each other, but they both have alternative versions where the segmentation and annotation is more similar to the existing monolingual UD treebanks for each language.

References

Lars Ahrenberg 2007. LinES: An English-Swedish Parallel Treebank. *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA, 2007)*.

Cristina Bosco, Simonetta Magni, Maria Simi 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank *7th Linguistic Annotation Workshop and Interoperability with Discourse.*

Janna Lipenkova and Milan Souček 2014. Converting Russian Dependency Treebank to Stanford Typed Dependencies Representation. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pages 143-147.

Marie-Catherine de Marneffe and Christopher D. Manning 2008 The Stanford typed dependencies representation. Workshop on Cross-framework and Cross-domain Parser Evaluation.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning 2014 Universal Stanford Dependencies: A cross-linguistic typology Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).

Ryan McDonald, Joakim Nivre, Yvonne Quimbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. *Proceedings of the 51st Annual Meeting of the ACL*, Sofia, Bulgaria, August 4-9 2013, pages 92-97.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12*, Istanbul, Turkey, May 23-25 2012.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter 2015. Universal Dependencies for Finnish. *Proceedings of the 20th Nordic Conference on Computational Linguistics*, Vilnius, Lithuania, May 12-13, 2015.

Pasi Tapanainen and Timo Järvinen 1997. A non-projective dependency parser. *Proceedings of the fifth conference on Applied Natural Language Processing*, pages 64-71.

Agić, Željko and Aranzabe, Maria Jesus and Atutxa, Aitziber and Bosco, Cristina and Choi, Jinho and de Marneffe, Marie-Catherine and Dozat, Timothy and Farkas, Richárd and Foster, Jennifer and Ginter, Filip and Goenaga, Iakes and Gojenola, Koldo and Goldberg, Yoav and Hajič, Jan and Johannsen, Anders Trærup and Kanerva, Jenna and Kuokkala, Juha and Laippala, Veronika and Lenci, Alessandro and Lindén, Krister and Ljubešić, Nikola and Lynn, Teresa and Manning, Christopher and Martínez, Héctor Alonso and McDonald, Ryan and Missilä, Anna and Montemagni, Simonetta and Nivre, Joakim and Nurmi, Hanna and Osenova, Petya and Petrov, Slav and Piitulainen, Jussi and Plank, Barbara and Prokopidis, Prokopis and Pyysalo, Sampo and Seeker, Wolfgang and Seraji, Mojgan and Silveira, Natalia and Simi, Maria and Simov, Kiril and Smith, Aaron and Tsarfaty, Reut and Vincze, Veronika and Zeman, Daniel Universal Dependencies 1.1. 2015. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/LRT-1478

Universal Dependencies Universal Dependencies home page. 2015. http://universaldependencies.github.io/docs/