Empirical lower bounds on alignment error rates in syntax-based machine translation

Anders Søgaard[*]
Center for Language Technology
University of Copenhagen
soegaard@hum.ku.dk

Jonas Kuhn[†]
Dpt. of Linguistics
University of Potsdam
kuhn@ling.uni-potsdam.de

Abstract

The empirical adequacy of synchronous context-free grammars of rank two (2-SCFGs) (Satta and Peserico, 2005), used in syntax-based machine translation systems such as Wu (1997), Zhang et al. (2006) and Chiang (2007), in terms of what alignments they induce, has been discussed in Wu (1997) and Wellington et al. (2006), but with a one-sided focus on so-called "inside-out alignments". Other alignment configurations that cannot be induced by 2-SCFGs are identified in this paper, and their frequencies across a wide collection of hand-aligned parallel corpora are examined. Empirical lower bounds on two measures of alignment error rate, i.e. the one introduced in Och and Ney (2000) and one where only complete translation units are considered, are derived for 2-SCFGs and related formalisms.

1 Introduction

Syntax-based approaches to machine translation typically use synchronous grammars to recognize or produce translation equivalents. The synchronous production rules are typically learned from alignment structures (Wu, 1997; Zhang and Gildea, 2004; Chiang, 2007) or from alignment structures and derivation trees for the source string (Yamada and Knight, 2001; Zhang and Gildea, 2004). They are also used for inducing alignments (Wu, 1997; Zhang and Gildea, 2004).

It is for all three reasons, i.e. translation, induction from alignment structures and induction of alignment structures, important that the synchronous grammars are expressive enough to induce all the alignment structures found in hand-aligned gold standard parallel corpora (Wellington et al., 2006). Such alignments are supposed to reflect the structure of translations, typically contain fewer errors and are used to evaluate automatically induced alignments.

In this paper it is shown that the synchronous grammars used in Wu (1997), Zhang et al. (2006) and Chiang (2007) are not expressive enough to do that. The synchronous grammars used in these systems are, formally, synchronous context-free grammars of rank two (2-SCFGs), or equivalently (normal form) inversion transduction grammars (ITGs).[1] The notion of *rank* is defined as the maximum number of constituents aligned by a production rule, i.e. the maximum number of distinct indeces. Our results will be extended to slight extensions of 2-SCFGs, incl. the extension of ITGs proposed by Zens and Ney (2003) (xITGs), synchronous tree substitution grammars of rank two (2-STSGs) (Eisner, 2003; Shieber, 2007), i.e. where tree pairs include at most two linked pairs of nonterminals, and synchronous tree-adjoining grammars of rank two

[*]This work was done while the first author was a Senior Researcher at the Dpt. of Linguistics, University of Potsdam, supported by the German Research Foundation in the Emmy Noether project *Ptolemaios* on grammar learning from parallel corpora; and while he was a Postdoctoral Researcher at the ISV Computational Linguistics Group, Copenhagen Business School, supported by the Danish Research Foundation in the project *Efficient syntax- and semantics-based machine translation*.

[†]The second author is supported by the German Research Foundation in the Emmy Noether project *Ptolemaios* on grammar learning from parallel corpora.

[1]2-SCFGs allow distinct LHS nonterminals, while ITGs do not; but for any 2-SCFG an equivalent ITG can be constructed by creating a cross-product of nonterminals from two sides.

Proceedings of SSST-3, Third Workshop on Syntax and Structure in Statistical Translation, pages 19–27,
Boulder, Colorado, June 2009.

(2-STAGs) (Shieber and Schabes, 1990; Harbusch and Poller, 1996; Nesson et al., 2008). The overall frequency of alignment structures that cannot be induced by these approaches is examined across a wide collection of hand-aligned parallel corpora. Empirical lower bounds on the coverage of the systems are derived from our results.

Our notion of an alignment structure is standard. Words can be aligned to multiple words. Unaligned nodes are permitted. Maximally connected subgraphs are called translation units. There is one more choice to make in the context of many-to-many alignments, namely whether the alignment relation is such that if $w_i|w'_k$ and $w_i|w'_l$, resp., are aligned, and $w_j|w'_k$ are aligned too, then $w_j|w'_l$ are also aligned. If so, the alignment structure is divided into complete translation units. Such alignment structures are therefore called *complete*; in Goutte et al. (2004), alignment structures with this property are said to be closed under transitivity. An alignment structure is simply written as a sequence of alignments, e.g. $\langle w_i|w'_k, w_i|w'_l, w_j|w'_k, w_j|w'_k\rangle$, or, alternatively, as sequences of (possibly discontinuous) translation units, e.g. $\langle w_i w_j|w'_k w'_l\rangle$.

A translation unit induced by a synchronous grammar is a set of terminals that are recognized or generated simultaneously. Consequently, synchronous grammars can only induce complete alignment structures (by transitivity of simultaneity).[2]

Syntax-based approaches to machine translations are commonly evaluated in terms of their alignment error rate (AER) on one or more parallel corpora (Och and Ney, 2000; Zhang and Gildea, 2004). The AER, in the case where all alignments are sure alignments, is

$$\text{AER} = 1 - \frac{2|S_A \cap G_A|}{|S_A| + |G_A|}$$

where G_A are the gold standard alignments, and S_A the alignments produced by the system.

AER has been criticized by Fraser and Marcu (2007). They show that AER does not penalize unequal precision and recall when a distinction between sure and possible alignments is made. Since no such distinction is assumed below, the classical definition is used.

We introduce also the notion of *translation unit error rate* (TUER), which is defined as

$$\text{TUER} = 1 - \frac{2|S_U \cap G_U|}{|S_U| + |G_U|}$$

where G_U are the translation units in the gold standard, and S_U the translation units produced by the system. In other words, what is measured is a system's ability to predict translation units relative to the Gold standard, not just its ability to predict alignments. If the system only gets part of a translation unit right, it is not rewarded.

In the context of many-to-many alignments, this measure may tell us more about translation quality than AER. Consider, for instance, the small children's book discourse in Danish:

(1) *Mads og Mette lægger tal*
Mads CONJ Mette put.FIN.PRES number.PL
sammen.
together
'Mads and Mette add numbers.'

(2) *Mads og Mette lægger tal*
Mads CONJ Mette put.FIN.PRES number.PL
sammen hver dag.
together every day
'Mads and Mette add numbers every day.'

(3) *Mads og Mette kan godt*
Mads CONJ Mette can.FIN.PRES good
lide at addere.
like.INF to add.INF
'Mads and Mette like to add.'

(4) *Mette spørger ofte:*
Mette ask.FIN.PRES often:
Skal vi addere
Shall.FIN.FUT/PRES PRON.PL.1 add.INF
sammen?
together
'Mette often asks: Do you want to add together?'

Say (1-4) and the English translations are a parallel corpus on which we would like to evaluate an aligner or a statistical machine translation system. Say also that the test corpus has been aligned. Let the first three sentences be our training data and (4) our test data.

[2] One of the hand-aligned parallel corpora used in our experiments, the one also used in Padó and Lapata (2006), includes incomplete alignment structures.

Note that the words *lægger . . . sammen* form a discontinuous translation unit ('add'). Say our aligner aligned only *sammen* and *add*, but not *lægger* and *add*. This would mean that the alignments or translations of *add* would most likely be associated with the following probabilities:

.66 (*add, sammen*)
.33 (*add, addere*)

which again means that our system is likely to arrive at the wrong alignment or translation in (4). Nevertheless these alignments are rewarded in AER. TUER, on the other hand, reflects the intuition that unless you get the entire translation unit it's better to get nothing at all.

The hand-aligned parallel corpora in our experiments come from the Copenhagen Dependency Treebank (Buch-Kromann, 2007), for five different language pairs, the German-English parallel corpus used in Padó and Lapata (2006), and the six parallel corpora of the first 100 sentences of Europarl (Koehn, 2005) for different language pairs documented in Graca et al. (2008). Consequently, our experiments include a total of 12 parallel corpora. The biggest parallel corpus consists of 4,729 sentence pairs; the smallest of 61 sentence pairs. The average size is 541 sentence pairs. The six parallel corpora documented in Graca et al. (2008) use sure and possible alignments; in our experiments, as already mentioned, the two types of alignments are treated alike.[3]

[3]The annotations of the parallel corpora differ in format and consistency. In fact the empirical lower bounds obtained below are lower bounds in two senses: (i) they are lower bounds on TUERs because TUERs may be significantly higher than the empirical lower bounds found here, and (ii) they are lower bounds in the sense that there may be hidden instances of the configurations in question in the parallel corpora. Most seriously, our search algorithms only sort alignments, but not their elements; instead they assume that their elements are listed in chronological order. Sometimes, but rarely, this is not the case. Consider, for instance, file 1497, line 12 in the Danish–Spanish parallel corpus in the Copenhagen Dependency Treebank:

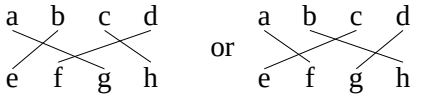
<align out="a56" type="" in="b30+b32+b8" outsign="af" insign="del de de"/>

This is a translation unit. The word in position 56 in the source string is aligned to the words in positions 8, 30 and 32 in the target string, but note that the target string words do not appear in chronological order. In some cases our algorithms take care of this; they do not, however, in general search all possible combinations of words and alignments, but rely on the linear order of the annotation. This was necessary to do relatively efficient queries. The effect, however, is that our results are lower than the actual frequencies in the parallel corpora. They are in this sense also lower bounds.

Sect. 2 discusses the frequency of inside-out alignments in our hand-aligned corpora, whereas Sect. 3 is about complex translation units. Sect. 4 briefly introduces formalisms for syntax-based machine translation, but some prior knowledge is assumed. Sect. 5 brings the three sections together and presents lower bounds on the coverage of the systems discussed in Sect. 4, obtained by inspection of the results in Sect. 2 and 3. Sect. 6 compares our results to related work, in particular Zens and Ney (2003).

2 Inside-out alignments

Wu (1997) identified so-called inside-out alignments, two alignment configurations that cannot be induced by binary synchronous context-free grammars; these alignment configurations, while infrequent in language pairs such as English–French (Cherry and Lin, 2006; Wellington et al., 2006), have been argued to be frequent in other language pairs, incl. English–Chinese (Wellington et al., 2006) and English–Spanish (Lepage and Denoual, 2005). While our main focus is on configurations that involve discontinuous translation units, the frequencies of inside-out alignments in our parallel corpora are also reported. Recall that inside-out alignments are of the form (or upside-down):



Our findings are summarized in Figure 1. Note that there is some variation across the corpora. The fact that there are no inside-out alignments in corpora 2–4 may be because annotators of these corpora have been very conservative, i.e. there are many unaligned nodes; the first corpus, which is also part of the Danish Dependency Treebank, also has very few inside-out alignments. It is not entirely clear to us if this has to do with the languages in question or the annotation guide lines (cf. Danish–Spanish).

In the Danish–Spanish corpus and in the English–German corpus the number of inside-out alignments is very high. This, to some extent, has to do with the number of words that are aligned to multiple words.

	Snt.	TUs	IO	IO-m	IO-m/Snt.
Danish–English:	4,729	110,511	28	4	0.001
Danish–German:	61	1,026	0	0	0
Danish–Italian:	181	2,182	0	0	0
Danish–Russian:	61	618	0	0	0
Danish–Spanish:	710	6,110	2,562	158	0.223
English–German	987	68,760	191,490	1,178	1.194
English–French:	100	937	2,651	80	0.800
English-Portuguese:	100	941	3,856	66	0.660
English–Spanish:	100	950	2,287	67	0.670
Portuguese–French:	100	915	3,643	84	0.840
Portuguese–Spanish:	100	991	1,194	58	0.580
Spanish–French	100	975	1,390	61	0.610

Figure 1: Frequency of inside-out alignments.

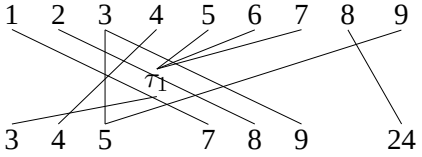
Say, in the case of English–German, each inside-out alignment is made out of eight two-word translation units. There are 1,178 inside-out alignment *modulo* translation units, i.e. when one or more inside-out alignments over the same eight translation units only count as one; this means that there would be $2^8 \times 1,178 : 301,568$ inside-out alignments in total. The actual number $(191,491)$ is smaller, but comparable.

The first example in the English–German corpus, from sentence 2, illustrates this point. The sentences are:

(5) Mr Jonckheer, I would like to thank you just as warmly for your report on the seventh survey on State aid in the European Union .

(6) Ebenso herzlich möchte ich Ihnen, Herr Jonckheer, für Ihren Bericht über den siebenten Bericht über staatliche Beihilfen in der Europäischen Union danken (24).

and the alignment structure is (commas count):



The aligned translation units are:[4]

⟨Mr|Herr⟩ ⟨Jonckheer|Jonckheer⟩
⟨,|Ihnen . . . ,⟩ ⟨I|ich⟩
⟨would like to|möchte⟩ ⟨thank|danken⟩
⟨you|Ihnen⟩

Note that the following sets of alignments make up distinct inside-out alignments *modulo* translation units:

$$\{\langle 1|7,4|4,8|24,9|5\rangle, \langle 2|8,4|4,8|24,9|5\rangle, \langle 3|9,4|4,8|24,9|5\rangle, \langle 1|7,5|3,8|24,9|5\rangle, \langle 2|8,5|3,8|24,9|5\rangle, \langle 3|9,5|3,8|24,9|5\rangle\}$$

The following sets of alignments in addition make up distinct inside-out alignments, but the new alignments $6|3$ and $7|3$ are from the same translation unit as $5|3$:

$$\{\langle 1|7,6|3,8|24,9|5\rangle, \langle 2|8,6|3,8|24,9|5\rangle, \langle 3|9,6|3,8|24,9|5\rangle, \langle 1|7,6|3,8|24,9|5\rangle, \langle 2|8,6|3,8|24,9|5\rangle, \langle 3|9,6|3,8|24,9|5\rangle\}$$

Consequently, the alignment of sentences (5) and (6) in the English–German parallel corpus contains 12 inside-out alignments, but only six inside-out alignments *modulo* translation units.

3 Cross-serial discontinuous translation units

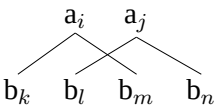
A discontinuous translation unit (DTU) is a translation unit where either the substring of source string words or the substring of target string words that occur in it, is discontinuous, i.e. there is a gap in it.

Since translation units are induced by simultaneous recognition, it is necessary for synchronous

[4] Note that the alignment $3|5$ is probably a mistake made by the annotator. It should, it seems, be $3|6$. Note also that this alignment is not involved in any of the inside-out alignments.

grammars to have rules that introduce multiple source side terminals and/or multiple target side terminals with at least one intervening nonterminal to induce DTUs. A DTU with multiple gaps in the same side is called a multigap DTU; it is easy to see that binary grammars cannot induce multigap DTUs with more than two gaps.

A sequence of DTUs is said to be *cross-serial* if it is of the following form (or upside-down):



Call any sequence of cross-serial DTUs a cross-serial DTU (CDTU). So a CDTU is an alignment configuration such that the source-side, resp. target-side, contains four tokens b_k, b_l, b_m, b_n such that (i) $b_k \prec b_l \prec b_m \prec b_n$, (ii) b_k and b_m belong to the same translation unit T, and b_l and b_n belong to the same translation unit T', and (iii) T and T' are distinct translation units. The inability of ITGs, xITGs and 2-STSGs to induce CDTUs follows from the observation that if b_k and b_m in the above are generated or recognized simultatenously in any of these formalisms, b_l and b_n cannot be generated or recognized simulaneously. This is a straight-forward consequence of the context-freeness of the component grammars.

The distinction between CDTUs and CDTUs *modulo* translation units (CDTU-ms) is again important. The number of CDTU-ms is the number of CDTUs such that all CDTUs differ by at most one translation unit. The English–German parallel corpus, for example, contains 15,717 CDTUs, but only 2,079 CDTU-ms. Since our evaluation measure is TUER, we only systematically counted the occurrences of CDTU-ms. In a few cases, the number of CDTUs was extracted too. In general, it was about eight times higher than the number of CDTU-ms.

Our findings are summarized in Figure 2. There is again variation, but the average ratio of CDTU-ms is 0.514, i.e. there is a CDTU-m in about every second aligned sentence pair.

4 Syntax-based machine translation

Syntax-directed translation schemas (SDTSs) were originally introduced by Culik (1966) and studied formally by Aho and Ullman (1972), who stressed the importance of using only binary SDTSs for efficiency reasons,[5] and later led to the development of a number of near-equivalent theories, incl. 2-SCFGs and (normal form) ITGs. Henceforth, we will refer to this class of near-equivalent theories as ITGs (see footnote 1). This also means that production rules have at most one source-side and one target-side terminal on the RHS (see below).

It is the ability of ITGs to induce alignments that is our main focus. Related work includes Wu (1997), Zens and Ney (2003) and Wellington et al. (2006). Our results will also be extended to xITGs, 2-STSGs and 2-STAGs. $\mathcal{O}(|G|n^6)$ time recognition algorithms are known for ITGs, xITGs and 2-STSGs. 2-STAGs ($\mathcal{O}(|G|n^{12})$) are more complex.

The production rules in ITGs are of the following form (Wu, 1997), with a notation similar to what is typically used for SDTSs and SCFGs in the right column:

$A \rightarrow [BC]$	$A \rightarrow \langle B^1C^2, B^1C^2 \rangle$
$A \rightarrow \langle BC \rangle$	$A \rightarrow \langle B^1C^2, C^2B^1 \rangle$
$A \rightarrow e \mid f$	$A \rightarrow \langle e, f \rangle$
$A \rightarrow e \mid \epsilon$	$A \rightarrow \langle e, \epsilon \rangle$
$A \rightarrow \epsilon \mid f$	$A \rightarrow \langle \epsilon, f \rangle$

It is important to note that RHSs of production rules have at most one source-side and one target-side terminal symbol. This prevents induction of multiword translation units in any straight-forward way. xITGs (Zens and Ney, 2003) in part solves this problem. All production rules in ITGs can be production rules in xITGs, but xITG production rules can also be of the following form:

$$A \quad \rightarrow \quad [e/f_1 A \epsilon/f_2] \mid \langle e/f_1 A \epsilon/f_2 \rangle$$

Note, however, that these production rules still do not enable double-sided DTUs, i.e. DTUs that translate into DTUs. Such, however, occur relatively frequently in hand-aligned parallel corpora, e.g. 148 times in the Danish–Spanish corpus.

There is no room for detailed introductions of the more complex formalisms, but briefly their differences can be summarized as follows:

The move from ITGs to 2-STSGs is relatively simple. All production rules in ITGs characterize

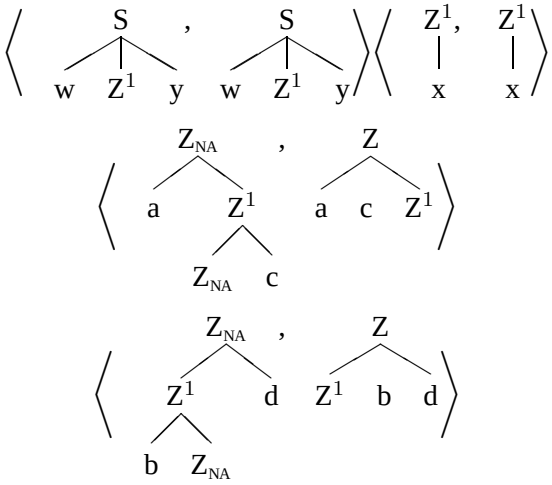
[5] The hierarchy of SDTSs of rank k is non-collapsing, and the recognition problem without a fixed rank is NP-hard (Aho and Ullman, 1972; Rambow and Satta, 1994). See Zhang et al. (2006) for an efficient binarization algorithm.

	Snt.	TUs	DTUs	DTUs/Snt.	CDTU-ms	CDTU-ms/Snt.
Danish–English:	4,729	110,511	1,801	0.381	6	0.001
Danish–German:	61	1,026	43	0.705	0	0
Danish–Italian:	181	2,182	63	0.348	1	0.006
Danish–Russian:	61	618	27	0.443	0	0
Danish–Spanish:	710	6,693	779	1.097	121	0.170
English–German	650	68,760	5,062	7.788	2,079	3.199
English–French:	100	937	95	0.950	38	0.380
English-Portuguese:	100	941	100	1.000	85	0.850
English–Spanish:	100	950	90	0.900	50	0.500
Portuguese–French:	100	915	77	0.770	27	0.270
Portuguese–Spanish:	100	991	80	0.800	55	0.550
Spanish–French	100	975	74	0.740	24	0.240

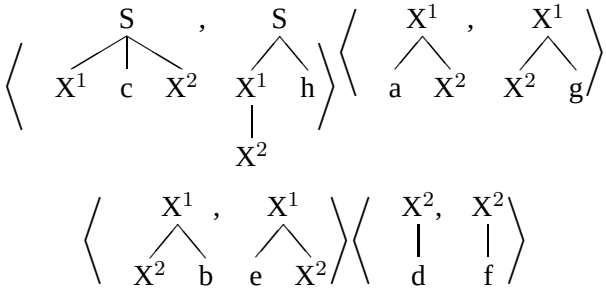
Figure 2: Frequency of cross-serial DTUs.

binary trees of depth 1. It is said that this is the domain of locality in ITGs. 2-STSGs extend the domain of locality to arbitrarily big trees. 2-STSGs are collections of ordered pairs of aligned trees with at most two pairs of linked nonterminals. The leaf nodes in the trees may be decorated by terminals or insertion slots where subtrees can be "plugged in". This is exactly what is meant by tree substitution. It is assumed that all terminals in a tree pair constitute a translation unit. There exists a $\mathcal{O}(|G|n^6)$ time parsing algorithm for 2-STSGs. 2-STSGs induce DTUs, double-sided DTUs and DTUs with at most two gaps, but *not* inside-out alignments, CDTUs and multigap DTUs with more than two gaps.

The substitution operation on elementary trees is supplied with an adjunction operation in 2-STAGs (Shieber and Schabes, 1990; Harbusch and Poller, 1996; Nesson et al., 2008). In adjunction, auxiliary trees, i.e. elementary trees with a designated leaf node labeled by a nonterminal identical to the nonterminal that labels the root node, extend the derived tree by expanding one of its nodes. If an auxiliary tree t, with a root node and a leaf node both labeled A, is adjoined at some node n also labeled A in a derived tree t', the subtree s' (of t') rooted at n is replaced by t, and s' is then inserted at the leaf node of t. In 2-STAGs, paired nodes across the source-side and target-side trees are simultaneously expanded by either substitution or adjunction. A $\mathcal{O}(|G|n^{12})$ parsing algorithm can be deviced for 2-STAGs using the techniques in Seki et al. (1991). The following 2-STAG translates Swiss-style cross-serial dependencies $\{wa^m b^n x c^m d^n y\}$ into $\{w(ac)^m x(bd)^n y\}$ and thus induces cross-serial DTUs whenever $m, n \geq 1$ (superscripts are pairings).



2-STAGs thus induce DTUs, double-sided DTUs, CDTUs, but not multigap DTUs with more than two gaps. 2-STAGs also induce inside-out alignments. Consider, for instance:



It is left for the reader to verify that this grammar induces the first of the two inside-out alignment configurations in Sect. 2.

5 Lower bounds on translation unit error rates

The ratio of inside-out alignments over TUs is a lower bound on the TUER for the binary versions of all the formalisms listed above, except 2-STAGs.

	IOs/TUs
Danish–English	0
Danish–German	0
Danish–Italian	0
Danish–Russian	0
Danish–Spanish	0.026
English–German	0.017
English–French	0.085
English–Portuguese	0.070
English–Spanish	0.070
Portuguese–French	0.092
Portuguese–Spanish	0.059
Spanish–French	0.063

For ITGs the ratio of DTUs over TUs is a lower bound on the TUER.

	DTUs/TUs
Danish–English	0.016
Danish–German	0.042
Danish–Italian	0.029
Danish–Russian	0.044
Danish–Spanish	0.121
English–German	0.074
English–French	0.101
English–Portuguese	0.106
English–Spanish	0.095
Portuguese–French	0.084
Portuguese–Spanish	0.081
Spanish–French	0.076

This is a considerable lower bound in itself, even for closely related languages such as Danish–German (4.2%) or Portuguese–Spanish (8.1%), which seems to have motivated research on extensions of ITGs (Zens and Ney, 2003). The ratio of CDTU-ms over TUs is a lower bound on the TUER for all the formalisms listed, except 2-STAGs:

	CDTU-ms/TUs
Danish–English	0
Danish–German	0
Danish–Italian	0.001
Danish–Russian	0
Danish–Spanish	0.018
English–German	0.030
English–French	0.041
English–Portuguese	0.090
English–Spanish	0.053
Portuguese–French	0.030
Portuguese–Spanish	0.056
Spanish–French	0.025

From these tables, empirical lower bounds on TUERs can be derived. ITGs, for instance, will have a TUER of at least $2.6\% + 12.1\% = 14.7\%$ for Danish–Spanish,[6] while 2-STSGs, ignoring problems caused by multigap DTUs with more than two gaps, will have a TUER of at least $7.0\% + 9.0\% = 16.0\%$ for English–Portuguese. Similarly lower bounds on AER for ITGs can be obtained by summing IOs/As, i.e. the number of inside-out alignments over the number of alignments, DTUs/As and CDTUs/As; for 2-STSGs, the lower bounds are given by IOs/As + CDTUs/As; and so on. Even 2-STAGs exclude alignments found in the data, namely multigap DTUs. The number of multigap DTUs (MDTUs) in the corpora documented in Graca et al. (2008) range from 3–11 (in a 100 sentences) with an average of 5.8. Exact results for each formalism that include double-sided DTUs and multigap DTUs will be included in a future publication, but it is clear to us that both configurations are less frequent than inside-out alignments and CDTUs. In the Danish–Spanish parallel corpus the number of DTUs with three or more gaps is 448 out of which 182 are CDTUs. In the English–German parallel corpus, the numbers are, resp., 2,529 and 996.

[6] It was recently suggested to us by a colleague that the lower bounds need not be additive. It is, theoretically, possible that the errors associated with CDTUs subsume some of the errors associated with inside-out alignments, i.e. that it is possible to remove one alignment or translation unit from the Gold standard alignment structure such that both the CDTU-ms count goes down by one, and the inside-out alignment count goes down by one. It is left for future work to estimate this bias, but it seems to us that such subsumptions will be infrequent.

6 Related work

Zens and Ney (2003) used GIZA++ to word-align the Verbmobil task (English and German) and the Canadian Hansards task (English and French) and tested the coverage of ITGs and xITGs, i.e. the ratio of the number of alignment configurations that could be induced by the theories and the sentences in the two tasks. The results are presented below:

	ITG	xITG
Verbmobil (G→E)	91.6%	96.5%
Verbmobil (E→G)	87.0%	96.9%
Can. Hansards (F→E)	81.3%	96.1%
Can. Hansards (E→F)	73.6%	95.6%

Note that the average differences in coverage between ITGs and xITGs for English-German (7.4%) and English–French (18.4%) are comparable to the DTUs/TUs ratios for English–German (7.4%), resp. English–French (10.1%) in our parallel corpora. Compare also the average error rate of xITGs for English and German (3.3%) and English and French (4.15%) to the CDTU-ms/TUs ratios for English–German (3.0%) and English–French (4.1%).

This data provides strong support that inside-out alignments and cross-serial DTUs are the main theoretical challenge for syntax-based machine translation; in addition, training is a major challenge (Zhang and Gildea, 2004). In real-life applications, AERs and TUERs will be significantly higher than the empirical lower bounds obtained here, e.g. 40% for Chinese–English in Zhang and Gildea (2004), but in principal future results should converge on them.

7 Discussion

In machine translation, as in all other branches of computer science, there is a trade-off between expressivity and complexity. The results presented here, namely that classes of alignment structures excluded by syntax-based translation systems, occur frequently in hand-aligned parallel corpora, could be taken to indicate that more expressive formalisms are needed. This at least seems to be the case to the extent alignment error rates are reasonable measures of the adequacy of syntax-based machine translation systems. On the other hand parsing complexities in syntax-based machine translation are very high already, i.e. $\mathcal{O}(|G|n^6)$ and higher. Consequently, it is not advisable to gain more expressivity at the expense of parsing complexity. This need not be necessary either, however. There are at least two other possibilities:

- Either the cake can be cut differently, i.e. to exclude other classes of alignment structures that occur less frequently. This idea has to the best of our knowledge not been explored in the context of syntax-based machine translation.
- It is also possible to design formalisms for syntax-based machine translation that induce all possible alignment structures and maintain a reasonable parsing complexity ($\mathcal{O}(|G|n^6)$), e.g. Søgaard (2008b); but as noted by Søgaard (2008a) the gain in expressivity is at the expense of the complexity of learning. Finally, it can be shown that there are no computable tight estimators for the probabilistic extension of the formalism introduced in Søgaard (2008b).[7]

8 Conclusion

It was shown how the frequency of certain classes of alignment structures induce empirical lower bounds on the alignment error rates that can be obtained with these systems. Some of these lower bounds are quite significant, e.g. 14.7% (TUER) for ITGs wrt. Danish–Spanish and 17.6% wrt. Portuguese-French. Slightly lower, but still significant, bounds exist for more complex formalisms such as 2-STSGs and 2-STAGs.

[7] Two other challenges for this type of approach are: (i) The use of intersection in Søgaard (2008b) to induce inside-out alignments and cross-serial DTUs seems to miss important generalizations; see Chiang (2004) for a similar point in the context of parsing. (ii) If the class of alignment structures is restricted in any natural way, i.e. to $1:1$ alignments, the problem whether there exists a possible alignment given two sentences and a grammar becomes NP-hard (Søgaard, 2009). NB: The undecidability of computing tight estimators was pointed out to us by Mark-Jan Nederhof (p.c.), but Alexander Clark (p.c.) and others have suggested that pseudo-tight estimators can be used in practice.

References

Alfred Aho and Jeffrey Ullman. 1972. *The theory of parsing, translation and compiling*. Prentice-Hall, London, England.

Matthias Buch-Kromann. 2007. Computing translation units and quantifying parallelism in parallel dependency treebanks. In *ACL'07, Linguistic Annotation Workshop*, pages 69–76.

Colin Cherry and Dekang Lin. 2006. A comparison of syntactically motivated word alignment spaces. In *EACL'06*, pages 145–152, Trento, Italy.

David Chiang. 2004. Uses and abuses of intersected languages. In *TAG+'04*, Vancouver, Canada.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

K. Culik. 1966. Well translatable languages and Algol-like languages. In T. Steel, editor, *Formal languages and description languages*, pages 76–85. N. Holland Press, Amsterdam, the Netherlands.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *ACL'03*, pages 205–208, Sapporo, Japan.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Cyril Goutte, Kenji Yamada, and Eric Gaussier. 2004. Aligning words using matrix factorisation. In *ACL'04*, pages 502–509, Barcelona, Spain.

Joao Graca, Joana Pardal, Luísa Coheur, and Diamantino Caseiro. 2008. Building a golden collection of parallel multi-language word alignments. In *LREC'08*, Marrakech, Morocco.

Karin Harbusch and Peter Poller. 1996. Structural translation with synchronous tree-adjoining grammars in Verbmobil. Technical Report Verbmobil 184, Universität Koblenz-Landau/DFKI GmbH, Koblenz, Germany.

Philipp Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *MT-Summit'05*, pages 79–86, Phuket, Thailand.

Yves Lepage and Etienne Denoual. 2005. Purest ever example-based machine translation. *Machine Translation*, 19(3–4):251–282.

Rebecca Nesson, Giorgio Satta, and Stuart Shieber. 2008. Optimal k-arization of synchronous tree-adjoining grammar. In *TAG+'08*, Tübingen, Germany.

Franz Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *COLING'00*, pages 1086–1090, Saarbrücken, Germany.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *ACL-COLING'06*, pages 1161–1168.

Owen Rambow and Giorgio Satta. 1994. A two-dimensional hierarchy for parallel rewriting systems. Technical report, University of Philadelphia, Philadelphia, Pennsylvania.

Giorgio Satta and Enoch Peserico. 2005. Some computational complexity results for synchronous context-free grammars. In *HLT-EMNLP'05*, pages 803–810, Vancouver, Canada.

Hiroyuki Seki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. 1991. On multiple context-free grammars. *Theoretical Computer Science*, 88(2):191–229.

Stuart Shieber and Yves Schabes. 1990. Synchronous tree-adjoining grammars. In *COLING'90*, pages 253–258, Helsinki, Finland.

Stuart Shieber. 2007. Probabilistic synchronous tree-adjoining grammars for machine translation. In *SSST'07*, pages 88–95, Rochester, New York.

Anders Søgaard. 2008a. Learning context-sensitive synchronous rules. In *EAMT'08*, pages 168–173, Hamburg, Germany.

Anders Søgaard. 2008b. Range concatenation grammars for translation. In *COLING'08*, pages 103–106, Manchester, England.

Anders Søgaard. 2009. The complexity of restricted alignment problems in two formalisms for syntax-based machine translation. In *SSST'09*, Boulder, Colorado. To appear.

Benjamin Wellington, Sonjia Waxmonsky, and Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *ACL'06*, pages 977–984, Sydney, Australia.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *ACL'01*, pages 531–538, Toulouse, France.

Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *ACL'03*, pages 144–151, Sapporo, Japan.

Hao Zhang and Daniel Gildea. 2004. Syntax-based alignment: supervised or unsupervised? In *COLING'04*, pages 418–424, Geneva, Switzerland.

Hao Zhang, Liang Huang, Daniel Gildea, and Kevin Knight. 2006. Synchronous binarization for machine translation. In *NAACL-HLT'06*, pages 256–263, New York, New York.