Synchronous Linear Context-Free Rewriting Systems for Machine Translation

Miriam Kaeshammer
University of Düsseldorf
Universitätsstraße 1
40225 Düsseldorf, Germany

Abstract

We propose synchronous linear context-free rewriting systems as an extension to synchronous context-free grammars in which synchronized non-terminals span $k \geq 1$ continuous blocks on each side of the bitext. Such discontinuous constituents are required for inducing certain alignment configurations that occur relatively frequently in manually annotated parallel corpora and that cannot be generated with less expressive grammar formalisms. As part of our investigations concerning the minimal k that is required for inducing manual alignments, we present a hierarchical aligner in form of a deduction system. We find that by restricting k to 2 on both sides, 100% of the data can be covered.

1 Introduction

The most prominent paradigms in statistical machine translation are phrase-based translation models (Koehn et al., 2003) and tree-based approaches using some form of a synchronous context-free grammar (SCFG) (Chiang, 2007; Zollmann and Venugopal, 2006; Hoang and Koehn, 2010), in particular inversion transduction grammar (ITG) (Wu, 1997). The rules of the translation models are usually learned from word aligned parallel corpora. Synchronous grammars also induce alignments between words in the bitext when simultaneously recognizing words via the application of a synchronous rule (Wu, 1997). Due to their central role, it is important that a synchronous grammar formalism is powerful enough to generate all alignment configurations that occur in hand-aligned parallel corpora that are taken to be a gold standard of translational equivalence (Wellington et al., 2006).

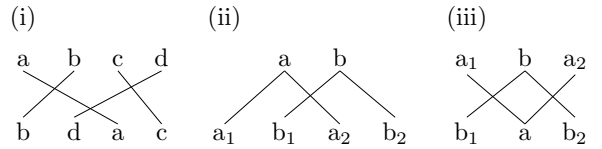

Figure 1: (i) inside-out alignment (Wu, 1997); (ii) cross-serial discontinuous translation unit (Søgaard and Kuhn, 2009); (iii) bonbon alignment (Simard et al., 2005)

The empirical adequacy of phrase-based and SCFG-based translation models has been put into question (Wellington et al., 2006; Søgaard and Kuhn, 2009; Søgaard and Wu, 2009; Søgaard, 2010) because they are unable to induce certain alignment configurations. In the alignments in Figure 1, the translation units *a*, *b*, *c*, and *d* cannot be independently generated by a binary SCFG. Due to a reordering component, phrase-based systems can handle (i), but neither (ii) nor (iii). Those phenomena however occur relatively frequently in hand-aligned parallel corpora. Wellington et al. (2006) found that complex structures such as inside-out alignments occur in 5% of English-Chinese sentence pairs and in the study of Søgaard and Kuhn (2009) between 1.6% (for Danish-English data) and 12.1% (for Danish-Spanish data) of all translation units are discontinuous, i.e. not derivable by ITGs in normal form.

As Wellington et al. (2006) already noted for inside-out alignments, *discontinuous constituents* are required for binary synchronous derivations of the alignment configurations under consideration. This is illustrated in Figure 2: the yields of $A_{\boxed{2}}$

Proceedings of the 7th Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 68–77,
Atlanta, Georgia, 13 June 2013.

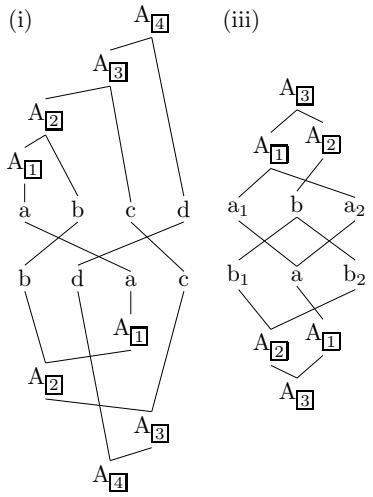

Figure 2: Synchronous derivations: co-indexed non-terminals are generated synchronously. Note that many other derivations that induce the same alignment structures are possible, but all of them involve at least one discontinuous constituent.

and $A_{\boxed{3}}$ in (i) are discontinuous on the target side, in (iii) the yield of $A_{\boxed{1}}$ is discontinuous on the source side and the yield of $A_{\boxed{2}}$ is discontinuous on the target side. We therefore propose to augment tree-based approaches such that they can account for discontinuous constituents in the source and/or target derivation. This implies going beyond the power of context-free grammars.

In the monolingual parsing community, linear context-free rewriting systems (LCFRS) have been established as an appropriate formalism for the modeling of discontinuous structure (Maier and Lichte, 2011; Kuhlmann and Satta, 2009). LCFRS is an extension of CFG, in which non-terminals can span $k \geq 1$ continuous blocks of a string. k is termed the *fan-out* of the non-terminal. If $k = 1$ for all non-terminals, the grammar is a CFG. Recent work shows that probabilistic data-driven parsing with LCFRS is indeed feasible and gives acceptable results (Maier, 2010; Evang and Kallmeyer, 2011; van Cranenburgh, 2012; Maier et al., 2012; Kallmeyer and Maier, 2013). It seems timely to transfer these findings to statistical machine translation.

In this work, we introduce the notion of synchronous LCFRS for translation and show how the alignments in Figure 1 are induced. Since the parsing complexity of LCFRS, and thus of synchronous LCFRS as well, depends directly on k, the number of blocks that a non-terminal in the grammar may span, an investigation concerning the empirically required k is carried out on manually aligned data. For this purpose, we present a parallel parser for an all-accepting synchronous LCFRS that is used to validate hierarchical alignments for a given k. This extends the work of Wellington et al. (2006) and Søgaard (2010) from a methodological point of view, as will be explained in Section 5. In particular, we will revise the results that Søgaard (2010) presented concerning the coverage of ITG. Our experiments furthermore include data sets that have not been used in previous similar studies.

2 Synchronous LCFRS for Translation

2.1 LCFRS

An LCFRS[1] (Vijay-Shanker et al., 1987; Weir, 1988) is a tuple $G = (N, T, V, P, S)$ where N is a finite set of non-terminals with a function dim: $N \to \mathbb{N}$ determining the *fan-out* of each $A \in N$; T and V are disjoint finite sets of terminals and variables; $S \in N$ is the start symbol with $dim(S) = 1$; and P is a finite set of rewriting rules

$$A(\alpha_1, \ldots, \alpha_{dim(A)}) \to A_1(X_1^{(1)}, \ldots, X_{dim(A_1)}^{(1)}) \cdots A_m(X_1^{(m)}, \ldots, X_{dim(A_m)}^{(m)})$$

where $A, A_1, \ldots, A_m \in N$, $X_j^{(i)} \in V$ for $1 \leq i \leq m$, $1 \leq j \leq dim(A_i)$ and $\alpha_i \in (T \cup V)^*$ for $1 \leq i \leq dim(A)$, for a *rank* $m \geq 0$. For all $r \in P$, every variable X in r occurs exactly once in the left-hand side (LHS) and exactly once in the right-hand side (RHS) of r. r describes how the yield of the LHS non-terminal is computed from the yields of the RHS non-terminals. The yield of S is the language of the grammar. Figure 3 shows a sample LCFRS with more explanations.

The *rank* of G is the maximal rank of any of its rules, and its *fan-out* is the maximal fan-out of any of its non-terminals. G is called a (u, v)-LCFRS if it has rank u and fan-out v.

2.2 Synchronous LCFRS

We define synchronous LCFRS (SLCFRS) in parallel to synchronous CFG, see for example Satta

[1] We use the syntax of simple range concatenation grammars (Boullier, 1998), a formalism that is equivalent to LCFRS.

$A(ab, cd) \to \varepsilon$	$\langle ab, cd \rangle$ in yield of A
$A(aXb, cYd) \to A(X, Y)$	if $\langle X, Y \rangle$ in yield of A, then also $\langle aXb, cYd \rangle$ in yield of A
$S(XY) \to A(X, Y)$	if $\langle X, Y \rangle$ in yield of A, then $\langle XY \rangle$ in yield of S

Figure 3: Sample LCFRS for $L = \{a^n b^n c^n d^n \mid n > 0\}$

and Peserico (2005). An SLCFRS is a tuple $G = (N_s, N_t, T_s, T_t, V_s, V_t, P, S_s, S_t)$ where N_s, T_s, V_s, S_s, resp. N_t, T_t, V_t, S_t are defined as for LCFRS. They denote the alphabets for the *source* and *target side* respectively. P is a finite set of synchronous rewriting rules $\langle r_s, r_t, \sim \rangle$ where r_s and r_t are LCFRS rewriting rules based on N_s, T_s, V_s and N_t, T_t, V_t respectively, and $\sim$ is a bijective mapping of the non-terminals in the RHS of r_s to the non-terminals in the RHS of r_t. This link relation is represented by co-indexation in the synchronous rules. During a derivation, the yields of two co-indexed non-terminals have to be explained from one synchronous rule. $\langle S_s, S_t \rangle$ is the start pair.

We call the tuple $(N_s, T_s, V_s, P_s, S_s)$ the *source side grammar* G_s and $(N_t, T_t, V_t, P_t, S_t)$ the *target side grammar* G_t where P_s is the set of all r_s in P and P_t is the set of all r_t in P. The *rank* u of G is the maximal rank of G_s and G_t, and the *fan-out* v of G is the sum of the fan-outs of G_s and G_t. We will sometimes write $v_{v_{G_s}|v_{G_t}}$ to make clear how the fan-out of G is distributed over the source and the target side. As in the monolingual case, a corresponding grammar G is called a (u, v)-SLCFRS.

As an example consider the rules in Figure 4. They translate cross-serial dependencies into nested ones. The rank of the corresponding grammar is 2 and its fan-out $4_{2|2}$.

Note that instead of defining an SLCFRS, one could also set the fan-out of each non-terminal in an LCFRS to ≥ 2, set $dim(S) = 2$, and formulate synchronization between the arguments of the non-terminals. The main disadvantage is that this requires $N_s = N_t$. Furthermore, this seems less perspicuous than SLCFRS when moving from SCFG to mild context-sensitivity. Generalized Multitext Grammar (Melamed et al., 2004) is another weakly equivalent grammar formalism.

In correspondence to ITG and normal-form ITG (NF-ITG) (Søgaard and Wu, 2009), we say an

$$\begin{aligned}
&\langle A(a, c) \to \varepsilon \quad , \; C(a, c) \to \varepsilon \rangle \\
&\langle B(b, d) \to \varepsilon \quad , \; D(bd) \to \varepsilon \rangle \\
&\langle A(aX, cZ) \to A_{\boxed{1}}(X, Z) \; , \; C(aX, Zc) \to C_{\boxed{1}}(X, Z) \rangle \\
&\langle B(bY, dU) \to B_{\boxed{1}}(Y, U) \; , \; D(bYd) \to D_{\boxed{1}}(Y) \rangle \\
&\langle S(XYZU) \to A_{\boxed{1}}(X, Z) B_{\boxed{2}}(Y, U) \, , \\
&\qquad\qquad S(XYZ) \to C_{\boxed{1}}(X, Z) D_{\boxed{2}}(Y) \rangle
\end{aligned}$$

Figure 4: Sample SLCFRS for $L = \{\langle a^n b^m c^n d^m, a^n b^m d^m c^n \rangle \mid n, m > 0\}$

SLCFRS G is in *normal form* if the following two conditions hold: (a) the rank of G is at most 2 and (b) for all $r \in P$ it holds that the LHS arguments of r_s and r_t contain either terminals or variables, but no mixture of both. The grammar in Figure 4 is not in normal form.

While ITGs constrain the order of the non-terminals in the RHS of the target side to be in the same or in the reverse order compared to the non-terminals in the RHS of the source side, we do not impose such ordering constraints (on the variables) for SLCFRS. However, it is obvious that a $(2, 2_{1|1})$-SLCFRS is equivalent to an ITG of rank 2 and that a $(2, 2_{1|1})$-SLCFRS in normal form is equivalent to a NF-ITG.

2.3 Alignment Capacity

A translation unit is a maximally connected subgraph of a given alignment structure. Typically this is the smallest unit from which translation models are learned. During a synchronous derivation, we interpret simultaneously recognized terminals as aligned (Wu, 1997). They thus correspond to a translation unit. We call the synchronous derivation tree a hierarchical alignment. Many-to-many alignments are interpreted conjunctively. This means that to induce a given translation unit, a grammar has to be able to generate the complete translation unit, and not just one of the corresponding word alignments. The last point has been argued for in Søgaard and Kuhn (2009).

SLCFRS are able to induce the alignment structures under consideration (in Figure 1). This is exemplified by the rules given in Figure 5.

Clearly, there exist many different possible hierarchical alignments for a given alignment structure. The underlying constraints for the grammars in Figure 5 are (a) each translation unit is represented by

(i)

$\langle A(a) \rightarrow \varepsilon \quad , A(a) \rightarrow \varepsilon \rangle$
$\langle A(Xb) \rightarrow A_{\boxed{1}}(X) \ , A(b, Y) \rightarrow A_{\boxed{1}}(Y) \rangle$
$\langle A(Xc) \rightarrow A_{\boxed{1}}(X) \ , A(Y_1, Y_2 c) \rightarrow A_{\boxed{1}}(Y_1, Y_2) \rangle$
$\langle A(Xd) \rightarrow A_{\boxed{1}}(X) \ , A(Y_1 d Y_2) \rightarrow A_{\boxed{1}}(Y_1, Y_2) \rangle$

(ii)

$\langle A(a) \rightarrow \varepsilon \quad , A(a_1, a_2) \rightarrow \varepsilon \rangle$
$\langle A(Xb) \rightarrow A_{\boxed{1}}(X) \ , A(Y_1 b_1 Y_2 b_2) \rightarrow A_{\boxed{1}}(Y_1, Y_2) \rangle$
or
$\langle A(b) \rightarrow \varepsilon \quad , A(b_1, b_2) \rightarrow \varepsilon \rangle$
$\langle A(aX) \rightarrow A_{\boxed{1}}(X) \ , A(a_1 Y_1 a_2 Y_2) \rightarrow A_{\boxed{1}}(Y_1, Y_2) \rangle$

(iii)

$\langle A(a_1, a_2) \rightarrow \varepsilon \quad , A(a) \rightarrow \varepsilon \rangle$
$\langle A(X_1 b X_2) \rightarrow A_{\boxed{1}}(X_1, X_2) \ , A(b_1 Y b_2) \rightarrow A_{\boxed{1}}(Y) \rangle$
or
$\langle A(b) \rightarrow \varepsilon \quad , A(b_1, b_2) \rightarrow \varepsilon \rangle$
$\langle A(a_1 X a_2) \rightarrow A_{\boxed{1}}(X) \ , A(Y_1 a Y_2) \rightarrow A_{\boxed{1}}(Y_1, Y_2) \rangle$

Figure 5: SLCFRS rules that induce the alignments in Figure 1. For (i) there are many other derivations possible, since there are $4!$ possibilities to combine the translation units in a binary way. The shown rules correspond to Figure 2(i).

exactly one rule and (b) each rule aligns exactly one translation unit and combines it with at most one already established synchronous constituent.

ITG and NF-ITG do not generate the same class of alignments (Søgaard and Wu, 2009). In parallel, a $(2, v)$-SLCFRS in normal form does not generate the same class of alignments as an unrestricted $(2, v)$-SLCFRS. Consider, for example, a discontinuous translation unit d with two gaps on the source side and a grammar G with fan-out $3_{2|1}$. G in normal form cannot induce d. In general, for generating x gaps, a fan-out of $x + 1$ is required. However, without the normal form requirement, G can possibly induce d with a rule that combines the terminals of d with the constituents that fill the gaps.

2.4 Parsing Complexity

LCFRS in normal form can be parsed in $\mathcal{O}(n^{3k})$ where k is the fan-out of the grammar (Seki et al., 1991). This result can be transferred to SLCFRS: An SLCFRS with fan-out v is essentially an LCFRS with fan-out $v + 1$. However, because of the start non-terminal S with $dim(S) = 2$, all non-terminals $A \in N$ with $dim(A) \geq 2$ and the special interpretation of the source/target side meaning that variables occur either on the source or target side but

$\langle T(\boldsymbol{\alpha_s}) \rightarrow \varepsilon \quad , T(\boldsymbol{\beta_t}) \rightarrow \varepsilon \rangle$
$\langle A(\boldsymbol{\alpha_1}) \rightarrow T_{\boxed{1}}(\boldsymbol{\alpha_1}) \quad , A(\boldsymbol{\beta_1}) \rightarrow T_{\boxed{1}}(\boldsymbol{\beta_1}) \rangle$
$\langle A(\boldsymbol{\alpha_1}) \rightarrow A_{\boxed{1}}(\boldsymbol{\alpha_2}) A_{\boxed{2}}(\boldsymbol{\alpha_3}) \ , A(\boldsymbol{\beta_1}) \rightarrow A_{\boxed{1}}(\boldsymbol{\beta_2}) A_{\boxed{2}}(\boldsymbol{\beta_3}) \rangle$

where $\boldsymbol{\alpha_s} \in (T_s{}^*)^{k_0}, \boldsymbol{\beta_t} \in (T_t{}^*)^{k'_0}, \boldsymbol{\alpha}_i \in (V_s{}^+)^{k_j}, \boldsymbol{\beta}_i \in (V_t{}^+)^{k'_j}$ for $0 < k_j \leq k_s, 0 < k'_j \leq k_t, 0 < i \leq 3, 0 \leq j \leq 3$

Figure 6: All-accepting SCLFRS in normal form with fan-out $v = k_s + k_t$

cannot change sides, no items that cross or involve the additional gap have to be built during parsing. Bitext parsing with SLCFRS in normal form can therefore also be performed in $\mathcal{O}(n^{3v})$ where $n = \max(n_s, n_t)$, or more specifically $\mathcal{O}(n_s^{3v_{G_s}} n_t^{3v_{G_t}})$ where n_s, n_t are the lengths of the source and target input strings respectively.

3 Empirical Investigation

Since parsing complexity with SLCFRS is determined by the fan-out v of the grammar, we conduct an investigation to find out which v would be required to fully cover the alignment configurations that occur in manually aligned parallel corpora.

3.1 Bottom-Up Hierarchical Aligner

Our study is based on *alignment validation* (Søgaard, 2010), i.e. we check whether an alignment structure can be generated by an all-accepting SLCFRS with a specific v. Such a grammar is depicted in Figure 6. Note in particular that it leaves open how to compose the yield of the LHS non-terminal from the two RHS constituents. To be able to use the grammar for parsing, one would have to spell out all combination possibilities.

Instead, we use the idea of a bottom-up hierarchical aligner (Wellington et al., 2006). It works very much like a synchronous parser, but the constraints for inferences are the word alignments and potentially other things, and not the rewriting rules of a grammar. Initial constituents are built from the word alignments, then constituents are combined with each other. The goal is to find a constituent that completely covers the input. In our case, the constraints for the hierarchical aligner come from the translation units, the fan-out $v_{k_s|k_t}$ of the simulated grammar and possibly a normal-form requirement.

We specify the hierarchical aligner in terms of a deduction system (Shieber et al., 1995). Deduc-

tion rules have the form $\frac{A_1 \dots A_m}{B} C$ where $A_1 \dots A_m$ and B are *items*, i.e. intermediate parsing results, and C is a list of conditions on $A_1 \dots A_m$ and B. The interpretation is that if $A_1 \dots A_m$ can be deduced and conditions C hold, then B can be deduced. Our items have the form $\langle [X_s, \boldsymbol{\rho}_s], [X_t, \boldsymbol{\rho}_t] \rangle$ where $X_s \in N_s$ and $X_t \in N_t$ of the simulated grammar. All-accepting grammars usually have only one non-terminal symbol, but we need a distinction between pre-terminal constituents T and general constituents A for simulating SLCFRS in normal form as well as the full class. $\boldsymbol{\rho}_s$ and $\boldsymbol{\rho}_t$ characterize the spans of the synchronous constituent on the source and target side respectively. We view them as bit vectors where $\boldsymbol{\rho}_s(i) = 1$ means that $\mathbf{s}_i$ is in the yield of X_s, and $\boldsymbol{\rho}_t(i') = 1$ that $\mathbf{t}_{i'}$ is in the yield of X_t. $\langle \mathbf{s}_{0 \dots n}, \mathbf{t}_{0 \dots n'} \rangle$ is the input sentence pair that is segmented into m disjoint translation units $\langle D_s^{(m)}, D_t^{(m)} \rangle$ based on the given word alignment structure. $D_s^{(m)}$ and $D_t^{(m)}$ are sets of word indices into $\mathbf{s}$ and $\mathbf{t}$ respectively. We furthermore specify some useful operations for bit vectors. The $\cup$ operator combines bit vectors of the same length to a new bit vector by an elementwise *or* operation, while the intersection $\cap$ of two bit vectors is the elementwise *and* operation. 0^l is a bit vector $\boldsymbol{\rho}$ such that $\boldsymbol{\rho}(i) = 0$ for all $0 \leq i \leq l$. The function $b(\boldsymbol{\rho})$ returns the number of blocks of $\boldsymbol{\rho}$, i.e. the number of continuous sequences of 1s in $\boldsymbol{\rho}$.

Figure 7 shows the deduction rules of the hierarchical aligner that simulate an all-accepting SLCFRS in normal form. *Scan* builds T items from translation units, *Unary* creates A items from T items, and *Binary* combines two A items to a larger A item. Via the side conditions, A items are only created if they respect the specified fan-out $v_{k_s|k_t}$ of the all-accepting grammar. If the hierarchical aligner finds an A item that spans $\langle \mathbf{s}, \mathbf{t} \rangle$, the alignment structure of $\langle \mathbf{s}, \mathbf{t} \rangle$ is valid, i.e. can be induced by an SLCFRS in normal form with fan-out $v_{k_s|k_t}$.

Since we are also interested in the empirical alignment capacity of SLCFRS without normal-form restriction, we present an extended deduction system in Figure 8. The additional rules lead to the simulation of an SLCFRS of rank 2 where terminals and variables can be combined in the arguments of the LHS non-terminals of the rewriting rules. Note in particular that the generation of T items is not constrained by a maximally allowed $v_{k_s|k_t}$.

For the computation of the items, we use standard chart parsing techniques, maintaining a chart and an agenda.

3.2 Data

We use manually aligned parallel corpora for our study.[2] Data sets that have already been previously used in similar experiments, e.g. in Wellington et al. (2006), Søgaard and Wu (2009), and Søgaard (2010), are those from Martin et al. (2005) for English-Romanian and English-Hindi, the English-French data from Mihalcea and Pedersen (2003), the Europarl data set described in Graça et al. (2008) for the six combinations of English, French, Portuguese and Spanish, the English-German Europarl data that was created for Padó and Lapata (2006), and data sets with Danish as the source language that are part of the Parole corpus of the Copenhagen Dependency Treebank (Buch-Kromann et al., 2009).

We furthermore perform our study on data sets that, to the best of our knowledge, have not been evaluated in a similar setting before. Those are English-Swedish gold alignments documented in Holmqvist and Ahrenberg (2011), the English-Inuktitut data used in Martin et al. (2005), more English-German data[3], the English-Spanish data set in Lambert et al. (2005) and English-Dutch alignments that are part of the Dutch Parallel Corpus (Macken, 2010). Characteristics about the data sets are presented in the last columns of Table 1.

3.3 Method

We apply the bottom-up hierarchical alignment algorithm in various configurations to each manually aligned sentence pair. If a goal item is found, the alignment structure can be induced with the formalism in question. We measure the number of sentence pairs for which a hierarchical alignment was reached over the total number of sentence pairs. Søgaard (2010) refers to this as *alignment reachability*, which is the inverse of *parse failure rate* (Wellington et al., 2006).

[2] Whenever there are sure (S) and possible (P) alignments annotated, we use both.

[3] By T. Schoenemann, from `http://user.phil-fak.uni-duesseldorf.de/~tosch/downloads.html`

Scan: $\dfrac{}{\langle [T, \boldsymbol{\rho}_s], [T, \boldsymbol{\rho}_t] \rangle}$ a translation unit $\langle D_s, D_t \rangle$
where $\boldsymbol{\rho}_s(i) = 1$ if $i \in D_s$, otherwise $\boldsymbol{\rho}_s(i) = 0$, and $\boldsymbol{\rho}_t(i') = 1$ if $i' \in D_t$, otherwise $\boldsymbol{\rho}_t(i') = 0$

Unary: $\dfrac{\langle [T, \boldsymbol{\rho}_s], [T, \boldsymbol{\rho}_t] \rangle}{\langle [A, \boldsymbol{\rho}_s], [A, \boldsymbol{\rho}_t] \rangle}$ $b(\boldsymbol{\rho}_s) \leq k_s, b(\boldsymbol{\rho}_t) \leq k_t$

Binary: $\dfrac{\langle [A, \boldsymbol{\rho}_s^1], [A, \boldsymbol{\rho}_t^1] \rangle, \langle [A, \boldsymbol{\rho}_s^2], [A, \boldsymbol{\rho}_t^2] \rangle}{\langle [A, \boldsymbol{\rho}_s^3], [A, \boldsymbol{\rho}_t^3] \rangle}$ $\boldsymbol{\rho}_s^1 \cap \boldsymbol{\rho}_s^2 = 0^n, \boldsymbol{\rho}_t^1 \cap \boldsymbol{\rho}_t^2 = 0^{n'}, b(\boldsymbol{\rho}_s^3) \leq k_s, b(\boldsymbol{\rho}_t^3) \leq k_t$
where $\boldsymbol{\rho}_s^3 = \boldsymbol{\rho}_s^1 \cup \boldsymbol{\rho}_s^2$ and $\boldsymbol{\rho}_t^3 = \boldsymbol{\rho}_t^1 \cup \boldsymbol{\rho}_t^2$

Goal: $\langle [A, \boldsymbol{\rho}_s], [A, \boldsymbol{\rho}_t] \rangle$
where $\boldsymbol{\rho}_s(i) = 1$ for all $0 \leq i \leq n$ and $\boldsymbol{\rho}_t(i') = 1$ for all $0 \leq i' \leq n'$

Figure 7: CYK deduction system for an all-accepting SLCFRS in normal form with fan-out $v_{k_s|k_t}$

UnaryMixed: $\dfrac{\langle [T, \boldsymbol{\rho}_s^T], [T, \boldsymbol{\rho}_t^T] \rangle, \langle [A, \boldsymbol{\rho}_s^A], [A, \boldsymbol{\rho}_t^A] \rangle}{\langle [A, \boldsymbol{\rho}_s], [A, \boldsymbol{\rho}_t] \rangle}$ $\boldsymbol{\rho}_s^T \cap \boldsymbol{\rho}_s^A = 0^n, \boldsymbol{\rho}_t^T \cap \boldsymbol{\rho}_t^A = 0^{n'}, b(\boldsymbol{\rho}_s) \leq k_s, b(\boldsymbol{\rho}_t) \leq k_t$
where $\boldsymbol{\rho}_s = \boldsymbol{\rho}_s^T \cup \boldsymbol{\rho}_s^A$ and $\boldsymbol{\rho}_t = \boldsymbol{\rho}_t^T \cup \boldsymbol{\rho}_t^A$

BinaryMixed: $\dfrac{\langle [T, \boldsymbol{\rho}_s^T], [T, \boldsymbol{\rho}_t^T] \rangle, \langle [A, \boldsymbol{\rho}_s^1], [A, \boldsymbol{\rho}_t^1] \rangle, \langle [A, \boldsymbol{\rho}_s^2], [A, \boldsymbol{\rho}_t^2] \rangle}{\langle [A, \boldsymbol{\rho}_s^3], [A, \boldsymbol{\rho}_t^3] \rangle}$ $\begin{array}{c} \boldsymbol{\rho}_s^T \cap \boldsymbol{\rho}_s^1 = 0^n, \boldsymbol{\rho}_s^1 \cap \boldsymbol{\rho}_s^2 = 0^n, \boldsymbol{\rho}_s^2 \cap \boldsymbol{\rho}_s^T = 0^n, \\ \boldsymbol{\rho}_t^T \cap \boldsymbol{\rho}_t^1 = 0^{n'}, \boldsymbol{\rho}_t^1 \cap \boldsymbol{\rho}_t^2 = 0^{n'}, \boldsymbol{\rho}_t^2 \cap \boldsymbol{\rho}_t^T = 0^{n'}, \\ b(\boldsymbol{\rho}_s^3) \leq k_s, b(\boldsymbol{\rho}_t^3) \leq k_t \end{array}$
where $\boldsymbol{\rho}_s^3 = \boldsymbol{\rho}_s^T \cup \boldsymbol{\rho}_s^1 \cup \boldsymbol{\rho}_s^2$ and $\boldsymbol{\rho}_t^3 = \boldsymbol{\rho}_t^T \cup \boldsymbol{\rho}_t^1 \cup \boldsymbol{\rho}_t^2$

Figure 8: Additional inference rules for the deduction system in Figure 7 for simulating an SLCFRS of rank 2 without normal form restriction.

		SLCFRS				Søgaard (2010)		**Data**			
		NF		$u = 2$							
		$v = 2_{1\|1}$ = NF-ITG	$v = 4_{2\|2}$	$v = 2_{1\|1}$ = ITG	$v = 4_{2\|2}$	NF-ITG	ITG	#SPs	min	med	max
Martin	en-ro (30)	45.07	**97.85**	95.07	**100.00**	-	-	447	2|2	20|19	96|94
	en-hi (40)	82.73	**100.00**	96.36	$^{1\|2}_{2\|1}$**100.00**	-	-	115	1|1	10|12	45|58
	en-iu (40)	40.66	**95.60**	100.00	**100.00**	-	-	100	10|3	26|10	79|26
Pado	en-de (15)	73.74	**100.00**	94.41	$^{1\|2}$**100.00**	38.97	45.13	987	5|5	24|23	40|40
Mihal.	en-fr	67.56	**98.88**	95.30	**100.00**	*76.98	*81.75	447	2|2	16|17	30|30
Graça	en-fr	73.00	**100.00**	95.00	$^{1\|2}$**100.00**	65.00	68.00	100	4|4	11|13	14|21
	en-pt	76.00	**100.00**	98.00	$^{1\|2}_{2\|1}$**100.00**	65.00	67.00	100	4|3	11|12	14|21
	en-es	82.00	**100.00**	96.00	$^{1\|2}_{2\|1}$**100.00**	73.00	74.00	100	4|4	11|11	14|24
	pt-fr	73.00	**97.00**	92.00	$^{1\|2}$**100.00**	63.00	63.00	100	3|4	12|13	21|21
	pt-es	90.00	**99.00**	99.00	$^{1\|2}_{2\|1}$**100.00**	80.00	81.00	100	3|4	12|11	21|24
	es-fr	74.00	**100.00**	91.00	$^{1\|2}$**100.00**	68.00	68.00	100	4|4	11|13	24|21
CDT	da-en (25)	72.90	**98.93**	97.80	**100.00**	-	-	5464	1|1	16|17	89|98
	da-de (25)	64.87	**98.42**	94.94	$^{1\|2}_{2\|1}$**100.00**	*47.62	*49.35	449	1|1	17|18	75|74
	da-es (25)	66.61	**97.68**	97.50	**100.00**	*30.68	*35.54	807	1|1	16|18	78|97
	da-it (25)	69.01	**97.65**	97.95	**100.00**	*60.00	*60.00	1514	1|1	16|19	78|268
Holmqv.	en-sv (30)	82.83	**99.78**	95.60	**100.00**	-	-	1164	1|1	21|19	40|40
Schoen.	en-de (40)	29.15	**94.74**	76.11	**100.00**	-	-	300	1|1	21|22	77|79
Lambert	en-es (40)	47.15	**97.83**	94.85	**100.00**	-	-	500	4|4	26|27	90|99
Macken	en-nl (30)	57.14	**98.86**	94.86	**100.00**	-	-	699	1|1	20|19	107|105

Table 1: Alignment reachability scores of our experiments and those of Søgaard (2010), plus characteristics of the data sets. The numbers in parentheses are the sentence length cut-offs used in our experiments. The results marked with * are not directly comparable to ours because different versions of the data sets were used.

3.4 Results

Table 1 shows the results. It confirmes that NF-ITG is not capable of generating the majority of alignment configurations. However, when allowing discontinuous constituents with maximally two blocks on each side ($v = 4_{2|2}$), NF-SLCFRS induces all alignments present in six of the data sets, and reaches scores > 97 for the other data sets, except two of them for which scores are still > 94.7.

For grammars without normal-form constraint, alignment reachability is generally higher. We tested grammars of rank 2 and found that over 90% of the sentence pairs in each data set can be induced without the necessity of discontinuous constituents (except data set *Schoen.*). Such grammars roughly correspond to successfully applied translation models, e.g. in Hiero (Chiang, 2007). Nevertheless, our experiments show that the gold alignments contain a proportion of structures that cannot be generated by ITGs. With a $(2, 4_{2|2})$-SLCFRS, all occurring alignment configurations are captured. For some data sets, a fan-out of 3 is enough to induce all alignments. This is indicated by ${}^{1|2}$ and ${}_{2|1}$.

Going back to grammars in normal form, the sentence pairs that cannot be induced with a grammar of fan-out $4_{2|2}$ all display translation units that require three (or very rarely four) blocks on at least the source or the target side. An interesting observation is that only the English-Inuktitut data can nevertheless be generated with fan-out 4, by distributing the allowed discontinuity unequally: with a NF-SLCFRS with fan-out $4_{3|1}$, the alignment reachability is 100. This is not surprising given the fact that Inuktitut is a polysynthetic language.

Previous results by Søgaard (2010) concerning the coverage of ITG and NF-ITG on hand-aligned data, repeated for convenience in Table 1, are much lower than ours and therefore present a highly distorted picture concerning the empirical need of discontinuous constituents. This is due to the fact that the implementation[4] used for the experiments handles unaligned words incorrectly. They are added deterministically to the first constituent that encounters them, which leads to false negatives as further explained in Figure 9. After fixing this issue, the same results as for NF-SLCFRS with $v = 2_{1|1}$ are obtained. Another problem of the implementation concerns discontinuous translation units. Søgaard's alignment validation returns *false* if the words in the gap are aligned, although such configurations are induced by unrestricted ITG, see Søgaard and Wu (2009, Section 3.2.1).

[4] `http://cst.dk/anders/itg-search.html`

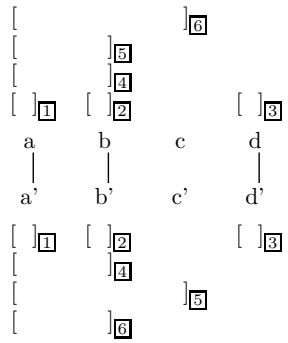

Figure 9: Synchronous ITG parse chart provided by the implementation from Søgaard (2010): *c* "belongs to" constituent 6 while *c'* "belongs to" constituent 5. When trying to combine 4 and 3, *c* and *c'* are not considered as unaligned because they are already part of a constituent, and neither 5 nor 6 can be combined with 3 without creating a discontinuous constituent. The algorithm cannot find a larger continuous constituent, the alignment validation therefore returns *false*. However, this simple alignment structure lies within the power of NF-ITG and ITG.

4 Discussion

Our experiments show that by moving from synchronous grammars with only continuous constituents to grammars that allow two blocks per constituent, (almost) all manual alignments can be generated, depending on whether the normal-form is enforced or not. Given the parsing complexity that comes with allowing discontinuities, this is a promising finding since it has already been shown for monolingual parsing that restricting the fan-out to 2 drastically reduces parsing times (Maier et al., 2012). In the future, we might also investigate whether refraining from ill-nested structures (Maier and Lichte, 2011) is a reasonable option for tree-based machine translation in order to reduce complexity (Gómez-Rodríguez et al., 2010).

Even though bitext parsing complexity for SLCFRS is prohibitively high, we expect that, given the techniques that have been developed for translation with SCFG, SLCFRS finds its application as a

translation model. In practice, only source side parsing is performed for translation and various pruning methods are applied to reduce the search space (e.g. in Chiang (2005), Yamada and Knight (2002) and many others).

It should also be mentioned that it is not clear yet how alignment reachability scores relate to machine translation quality and evaluation. We can nevertheless infer from the presented results that what is considered as translationally equivalent by the annotators of the data sets and their guidelines is beyond the search space of SCFG. A supplementary study could furthermore investigate translation unit error rates (Søgaard and Kuhn, 2009) for the data sets, under the assumption of a hierarchical SLCFRS alignment with a specific fan-out.

5 Related Work

Our empirical investigation extends previous studies, and thus provides new insights. Both Wellington et al. (2006) and Søgaard (2010) use a bottom-up hierarchical alignment algorithm with the goal of investigating the alignment complexity of manually aligned parallel corpora. Søgaard (2010) is however only interested in the alignment reachability of ITG and NF-ITG, and nothing beyond. We have furthermore revealed that the presented results underestimate the alignment capacity of ITG and NF-ITG.

The study of Wellington et al. (2006) is very similar to ours in that the number of blocks in discontinuous constituents that are required for hierarchical alignment are investigated. The word alignments are however treated disjunctively, which means that in the case of n-to-m alignments with $n, m \geq 1$, it is enough to induce one of the involved alignments. With this methodology a large class of discontinuities we are interested in, e.g. cross-serial discontinuous translation units, is ignored. The failure rates they present are therefore much lower than ours. Wellington et al. (2006) also show that when constraining synchronous derivations by monolingual syntactic parse trees on the source and/or target side, allowing discontinuous constituents becomes even more important for inducing gold alignments.

We are of course not the first to propose a translation model that is expressive enough to induce the alignments in question in Figure 1. Following up on a translation model proposed by Simard et al. (2005), Galley and Manning (2010) extend the phrase-based approach in that they allow for discontinuous phrase pairs. Their system outperforms a phrase-based system and a system based on SCFG of rank 2. In a way, our proposal to use SLCFRS is the syntax-based counterpart to their approach. Methods to integrate linguistic constituency information into the so far only formal tree-based approach can be directly transferred from the SCFG-based approaches to SLCFRS. In constrast, it is not obvious how to include such information into the phrase-based systems.

Søgaard (2008) proposes to use an even more expressive formalism than LCFRS, namely range concatenation grammar, and to exploit its ability to copy substrings during the derivation. The downsides of this approach are already mentioned in Søgaard and Kuhn (2009); for example, no tight probability estimation is possible for such a grammar.

The necessity of going towards mildly context-sensitive formalisms for translation modeling has also been advocated by Melamed (Melamed et al., 2004; Melamed, 2004). This step was however not motivated by the induction of specific complex translation units, but rather by the general observation that discontinuous constituents are necessary for synchronous derivations using linguistically motivated grammars. Discontinuous constituents also emerge when binarizing synchronous grammars of continuous yields with rank ≥ 4 (Melamed, 2003; Rambow and Satta, 1999).

6 Conclusion

Motivated by the finding that synchronous CFG cannot induce certain alignment configurations, we suggest to use synchronous LCFRS instead, which allows for discontinuities. Even though our empirical investigation shows that with exclusively continuous derivations more manual alignments can be captured than previously reported, there are still many aligned sentence pairs that can only be generated when setting the fan-out of the translation grammar to > 2. It remains to determine how such more accurate and more expressive models relate to translation quality.

Acknowledgments

I would like to thank Laura Kallmeyer and Wolfgang Maier for discussions and comments, the reviewers for their suggestions, Anders Søgaard for discussions concerning ITG and his implementation, and Zdeněk Žabokrtský for helping with the CDT data. This research was funded by the German Research Foundation as part of the project *Grammar Formalisms beyond Context-Free Grammars and their use for Machine Learning Tasks*.

References

Pierre Boullier. 1998. Proposal for a Natural Language Processing syntactic backbone. Technical Report 3342, INRIA.

Matthias Buch-Kromann, Iørn Korzen, and Henrik Høeg Müller. 2009. Uncovering the lost structure of translations with parallel treebanks. *Copenhagen Studies in Language*, 38:199–224.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270.

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

Kilian Evang and Laura Kallmeyer. 2011. PLCFRS parsing of English discontinuous constituents. In *Proceedings of the 12th International Conference on Parsing Technologies (IWPT)*, pages 104–116.

Michel Galley and Christopher D. Manning. 2010. Accurate non-hierarchical phrase-based translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 966–974.

Carlos Gómez-Rodríguez, Marco Kuhlmann, and Giorgio Satta. 2010. Efficient parsing of well-nested Linear Context-Free Rewriting Systems. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 276–284.

João de Almeida Varelas Graça, Joana Paulo Pardal, Luísa Coheur, and Diamantino António Caseiro. 2008. Building a golden collection of parallel multilanguage word alignment. In *The 6th International Conference on Language Resources and Evaluation (LREC08)*.

Hieu Hoang and Philipp Koehn. 2010. Improved translation with source syntax labels. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 409–417. Association for Computational Linguistics.

Maria Holmqvist and Lars Ahrenberg. 2011. A gold standard for English–Swedish word alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011*, pages 106–113.

Laura Kallmeyer and Wolfgang Maier. 2013. Data-driven parsing using Probabilistic Linear Context-Free Rewriting Systems. *Computational Linguistics*, 39(1). Accepted for publication.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54. Association for Computational Linguistics.

Marco Kuhlmann and Giorgio Satta. 2009. Treebank grammar techniques for non-projective dependency parsing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 478–486.

Patrik Lambert, Adrià Gispert, Rafael Banchs, and José B. Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39:267–285.

Lieve Macken. 2010. An annotation scheme and gold standard for Dutch-English word alignment. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Wolfgang Maier and Timm Lichte. 2011. Characterizing discontinuity in constituent treebanks. In *Formal Grammer 2009, Revised Selected Papers*, volume 5591 of *LNAI*. Springer.

Wolfgang Maier, Miriam Kaeshammer, and Laura Kallmeyer. 2012. Data-driven PLCFRS parsing revisited: Restricting the fan-out to two. In *Proceedings of the Eleventh International Conference on Tree Adjoining Grammars and Related Formalisms (TAG+11)*.

Wolfgang Maier. 2010. Direct parsing of discontinuous constituents in German. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*.

Joel Martin, Rada Mihalca, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *ACL Workshop on Building and Using Parallel Texts*.

I. Dan Melamed, Giorgio Satta, and Benjamin Wellington. 2004. Generalized multitext grammars. In *Proceedings of the 42nd Annual Conference of the Association for Computational Linguistics (ACL)*.

I. Dan Melamed. 2003. Multitext grammars and synchronous parsers. In *Proceedings of the 2003 Con-*

ference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 79–86.

I. Dan Melamed. 2004. Statistical machine translation by parsing. In *Proceedings of the 42nd Annual Conference of the Association for Computational Linguistics (ACL)*.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 1161–1168.

Owen Rambow and Giorgio Satta. 1999. Independent parallelism in finite copying parallel rewriting systems. *Theoretical Computer Science*, 223(1-2):87–120.

Giorgio Satta and Enoch Peserico. 2005. Some computational complexity results for synchronous context-free grammars. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 803–810.

Hiroyuki Seki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. 1991. On Multiple Context-Free Grammars. *Theoretical Computer Science*, 88(2):191–229.

Stuart M. Shieber, Yves Schabes, and Fernando C. N. Pereira. 1995. Principles and implementation of deductive parsing. *Journal of Logic Programming*, 24(1&2):3–36.

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with non-contiguous phrases. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 755–762.

Anders Søgaard and Jonas Kuhn. 2009. Empirical lower bounds on alignment error rates in syntax-based machine translation. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST '09)*. Association for Computational Linguistics.

Anders Søgaard and Dekai Wu. 2009. Empirical lower bounds on translation unit error rate for the full class of inversion transduction grammars. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 33–36.

Anders Søgaard. 2008. Range concatenation grammars for translation. In *Proceedings of Coling 2008: Companion volume: Posters*.

Anders Søgaard. 2010. Can inversion transduction grammars generate hand alignments? In *Proccedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT)*.

Andreas van Cranenburgh. 2012. Efficient parsing with Linear Context-Free Rewriting Systems. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

K. Vijay-Shanker, David Weir, and Aravind K. Joshi. 1987. Characterizing structural descriptions used by various formalisms. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*.

David Weir. 1988. *Characterizing Mildly Context-Sensitive Grammar Formalisms*. Ph.D. thesis, University of Pennsylviania, Philadelphia, PA.

Benjamin Wellington, Sonjia Waxmonsky, and I. Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 977–984.

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

Kenji Yamada and Kevin Knight. 2002. A decoder for syntax-based statistical MT. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation, HLT/NAACL*.