A Formal Characterization of Parsing Word Alignments by Synchronous Grammars with Empirical Evidence to the ITG Hypothesis

Gideon Maillette de Buy Wenniger*
University of Amsterdam
gemdbw@gmail.com

Khalil Sima'an*
University of Amsterdam
k.simaan@uva.nl

Abstract

Deciding whether a synchronous grammar formalism generates a given word alignment (the *alignment coverage problem*) depends on finding an adequate instance grammar and then using it to parse the word alignment. *But what does it mean to parse a word alignment by a synchronous grammar?* This is formally undefined until we define an unambiguous mapping between grammatical derivations and word-level alignments. This paper proposes an initial, formal characterization of alignment coverage as intersecting two *partially ordered sets (graphs)* of translation equivalence units, one derived by a grammar instance and another defined by the word alignment. As a first sanity check, we report extensive coverage results for ITG on automatic and manual alignments. Even for the ITG formalism, our formal characterization makes explicit many algorithmic choices often left underspecified in earlier work.

1 Introduction

The training data used by current statistical machine translation (SMT) models consists of source and target sentence pairs aligned together at the word level (*word alignments*). For the hierarchical and syntactically-enriched SMT models, e.g., (Chiang, 2007; Zollmann and Venugopal, 2006), this training data is used for extracting *statistically weighted Synchronous Context-Free Grammars (SCFGs)*. Formally speaking, a synchronous grammar defines a set of (source-target) sentence pairs derived synchronously by the grammar. Contrary to common belief, however, a synchronous grammar (see e.g., (Chiang, 2005; Satta and Peserico, 2005)) does not accept (or parse) word alignments. This is because a synchronous derivation generates a tree pair with a bijective binary relation (links) between their *nonterminal* nodes. For deciding whether a given word alignment is generated/accepted by a given synchronous grammar, it is necessary to *interpret* the synchronous derivations down to the lexical level. However, it is formally defined yet how to unambiguously interpret the synchronous derivations of a synchronous grammar as word alignments. One major difficulty is that synchronous productions, in their most general form, may contain *unaligned* terminal sequences. Consider, for instance, the relatively non-complex synchronous production

$$\langle X \rightarrow \alpha\, X^{(1)}\, \beta\, X^{(2)}\, \gamma\, X^{(3)},\ X \rightarrow \sigma\, X^{(2)}\, \tau\, X^{(1)}\, \mu\, X^{(3)} \rangle$$

where superscript (i) stands for aligned instances of nonterminal X and all Greek symbols stand for arbitrary non-empty terminals sequences. Given a word aligned sentence pair it is necessary to bind the terminal sequence by alignments consistent with the given word alignment, and then parse the word alignment with the thus enriched grammar rules. This is not complex if we assume that each of the source terminal sequences is contiguously aligned with a target contiguous sequence, but difficult if we assume arbitrary alignments, including many-to-one and non-contiguously aligned chunks.

One important goal of this paper is to propose a formal characterization of what it means to synchronously parse a word alignment. Our formal characterization is borrowed from the "parsing as intersection" paradigm, e.g., (Bar-Hillel et al., 1964; Lang, 1988; van Noord, 1995; Nederhof and Satta,

* Institute for Logic, Language and Computation.

Proceedings of the 7th Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 58–67,
Atlanta, Georgia, 13 June 2013.

2004). Conceptually, our characterization makes use of three algorithms. Firstly, parse the *unaligned* sentence pair with the synchronous grammar to obtain a set of synchronous derivations, i.e., trees. Secondly, interpret a word alignment as generating a set of synchronous trees representing the recursive translation equivalence relations of interest[1] perceived in the word alignment. And finally, *intersect* the sets of nodes in the two sets of synchronous trees to check whether the grammar can generate (parts of) the word alignment. The formal detail of each of these three steps is provided in sections 3 to 5.

We think that alignment parsing is relevant for current research because it highlights the difference between alignments in training data and alignments accepted by a synchronous grammar (learned from data). This is useful for literature on learning from word aligned parallel corpora (e.g., (Zens and Ney, 2003; DeNero et al., 2006; Blunsom et al., 2009; Cohn and Blunsom, 2009; Riesa and Marcu, 2010; Mylonakis and Sima'an, 2011; Haghighi et al., 2009; McCarley et al., 2011)). A theoretical, formalized characterization of the alignment parsing problem is likely to improve the choices made in empirical work as well. We exemplify our claims by providing yet another empirical study of the stability of the ITG hypothesis. Our study highlights some of the technical choices left implicit in preceding work as explained in the next section.

2 First application to the ITG hypothesis

A grammar *formalism* is a whole set/family of synchronous grammars. For example, ITG (Wu, 1997) defines a family of *inversion-transduction grammars* differing among them in the exact set of synchronous productions, terminals and non-terminals. Given a synchronous grammar *formalism* and an input word alignment, a relevant theoretical question is *whether there exists an instance synchronous grammar* that generates the word alignment exactly. We will refer to this question as the *alignment coverage* problem. In this paper we propose an approach to the alignment coverage problem using the three-step solution proposed above for parsing word alignments by arbitrary synchronous grammars.

Most current use of synchronous grammars is limited to a subclass using a pair of nonterminals, e.g., (Chiang, 2007; Zollmann and Venugopal, 2006; Mylonakis and Sima'an, 2011), thereby remaining within the confines of the ITG formalism (Wu, 1997). On the one hand, this is because of computational complexity reasons. On the other, this choice relies on existing empirical evidence of what we will call the "ITG hypothesis", freely rephrased as follows: the ITG formalism is sufficient for representing a major percentage of reorderings in translation data in general.

Although checking whether a word alignment can be generated by ITG is far simpler than for arbitrary synchronous grammars, there is a striking variation in the approaches taken in the existing literature, e.g., (Zens and Ney, 2003; Wellington et al., 2006; Søgaard and Wu, 2009; Carpuat and Wu, 2007; Søgaard and Kuhn, 2009; Søgaard, 2010). Søgaard and Wu (Søgaard and Wu, 2009) observe justifiably that the literature studying the ITG alignment coverage makes conflicting choices in method and data, and reports significantly diverging alignment coverage scores. We hypothesize here that the major conflicting choices in method (what to count and how to parse) are likely due to the absence of a well-understood, formalized method for parsing word alignments even under ITG. In this paper we apply our formal approach to the ITG case, contributing new empirical evidence concerning the ITG hypothesis.

For our empirical study we exemplify our approach by detailing an algorithm dedicated to ITG in Normal-Form (NF-ITG). While our algorithm is in essence equivalent to existing algorithms for checking binarizability of permutations, e.g.,(Wu, 1997; Huang et al., 2009), the formal foundations preceding it concern nailing down the choices made in parsing arbitrary word alignments, as opposed to (bijective) permutations. The formalization is our way to resolve some of the major points of differences in existing literature.

We report new coverage results for ITG parsing of manual as well as automatic alignments, showing the contrast between the two kinds. While the latter seems built for phrase extraction, trading-off precision for recall, the former is heavily marked with id-

[1]The translation equivalence relations of interest may vary in kind as we will exemplify later. The known phrase pairs are merely one possible kind.

iomatic expressions. Our coverage results make explicit a relevant dilemma. To hierarchically parse the current automatic word alignments *exactly*, we will need more general synchronous reordering mechanisms than ITG, with increased risk of exponential parsing algorithms (Wu, 1997; Satta and Peserico, 2005). But if we abandon these word alignments, we will face the exponential problem of learning reordering arbitrary permutations, cf. (Tromble and Eisner, 2009). Our results also exhibit the importance of explicitly defining the units of translation equivalence when studying (ITG) coverage of word alignments. The more complex the choice of translation equivalence relations, the more difficult it is to parse the word alignments.

3 Translation equivalence in MT

In (Koehn et al., 2003), a translation equivalence unit (TEU) is a *phrase pair*: a pair of contiguous substrings of the source and target sentences such that the words on the one side align only with words on the other side (formal definitions next). The hierarchical phrase pairs (Chiang, 2005; Chiang, 2007) are extracted by replacing one or more sub-phrase pairs, that are contained within a phrase pair, by pairs of linked variables. This defines a subsumption relation between hierarchical phrase pairs (Zhang et al., 2008). Actual systems, e.g., (Koehn et al., 2003; Chiang, 2007) set an upperbound on length or the number of variables in the synchronous productions. For the purposes of our theoretical study, these practical limitations are irrelevant.

We give two definitions of translation equivalence for word alignments.[2] The first one makes no assumptions about the contiguity of TEUs, while the second does require them to be contiguous substrings on both sides (i.e., phrase pairs).

As usual, $\mathbf{s} = s_1 ... s_m$ and $\mathbf{t} = t_1 ... t_n$ are source and target sentences respectively. Let $\mathbf{s}_\sigma$ be the source word at position σ in $\mathbf{s}$ and $\mathbf{t}_\tau$ be the target word at position τ in $\mathbf{t}$. An alignment link $a \in \mathbf{a}$ in a word alignment $\mathbf{a}$ is a pair of positions $\langle \sigma, \tau \rangle$ such that $1 \leq \sigma \leq m$ and $1 \leq \tau \leq n$. For the sake of brevity, we will often talk about alignments without explicitly mentioning the associated source and target words, knowing that these can be readily obtained from the pair of positions and the sentence pair $\langle \mathbf{s}, \mathbf{t} \rangle$. Given a subset $\mathbf{a}' \subseteq \mathbf{a}$ we define $words_{\mathbf{s}}(\mathbf{a}') = \{\mathbf{s}_\sigma \mid \exists X : \langle \sigma, X \rangle \in \mathbf{a}'\}$ and $words_{\mathbf{t}}(\mathbf{a}') = \{\mathbf{t}_\tau \mid \exists X : \langle X, \tau \rangle \in \mathbf{a}'\}$.

Now we consider triples $(\mathbf{s}', \mathbf{t}', \mathbf{a}')$ such that $\mathbf{a}' \subseteq \mathbf{a}$, $\mathbf{s}' = words_{\mathbf{s}}(\mathbf{a}')$ and $\mathbf{t}' = words_{\mathbf{t}}(\mathbf{a}')$. We define the *translation equivalence units (TEUs)* in the set $\mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ as follows:

Definition 3.1 $(\mathbf{s}', \mathbf{t}', \mathbf{a}') \in \mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ *iff* $\langle \sigma, \tau \rangle \in \mathbf{a}'$ $\Rightarrow$ *(for all X, if* $\langle \sigma, X \rangle \in \mathbf{a}$ *then* $\langle \sigma, X \rangle \in \mathbf{a}'$*)* $\wedge$ *(for all X, if* $\langle X, \tau \rangle \in \mathbf{a}$ *then* $\langle X, \tau \rangle \in \mathbf{a}'$*)*

In other words, if some alignment involving source position σ or τ is included in $\mathbf{a}'$, then all alignments in $\mathbf{a}$ containing that position are in $\mathbf{a}'$ as well. This definition allows a variety of complex word alignments such as the so-called *Cross-serial Discontiguous Translation Units* and *Bonbons* (Søgaard and Wu, 2009).

We also define the subsumption relation (partial order) $<_{\mathbf{a}}$ as follows:

Definition 3.2 *A TEU* $u_2 = (\mathbf{s_2}, \mathbf{t_2}, \mathbf{a_2})$ *subsumes* $(<_{\mathbf{a}})$ *a TEU* $u_1 = (\mathbf{s_1}, \mathbf{t_1}, \mathbf{a_1})$ *iff* $\mathbf{a_1} \subset \mathbf{a_2}$. *The subsumption order will be represented by* $u_1 <_{\mathbf{a}} u_2$.

Based on the subsumption relation we can partition $\mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ into two disjoint sets : atomic $\mathbf{TE_{Atom}}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ and composed $\mathbf{TE_{Comp}}(\mathbf{s}, \mathbf{t}, \mathbf{a})$.

Definition 3.3 $u_1 \in \mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ *is* atomic *iff* $\nexists\, u_2 \in \mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ *such that* $(u_2 <_{\mathbf{a}} u_1)$.

Now the set $\mathbf{TE_{Atom}}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ is simply the set of all atomic translation equivalents, and the set of composed translation equivalents $\mathbf{TE_{Comp}}(\mathbf{s}, \mathbf{t}, \mathbf{a}) = (\mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a}) \setminus \mathbf{TE_{Atom}}(\mathbf{s}, \mathbf{t}, \mathbf{a}))$.

Based on the general definition of translation equivalence, we can now give a more restricted definition that allows only contiguous translation equivalents (phrase pairs):

Definition 3.4 $(\mathbf{s}', \mathbf{t}', \mathbf{a}')$ *constitutes a* contiguous translation equivalent *iff:*

1. $(\mathbf{s}', \mathbf{t}', \mathbf{a}') \in \mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ *and*

[2] Unaligned words tend to complicate the formalization unnecessarily. As usual we also require that unaligned words must first be grouped with aligned words adjacent to them before translation equivalence is defined for an alignment. This standard strategy allows us to informally discuss unaligned words in the following without loss of generality.

2. *Both* $\mathbf{s}'$ *and* $\mathbf{t}'$ *are contiguous substrings of* $\mathbf{s}$ *and* $\mathbf{t}'$ *respectively.*

This set of translation equivalents is the unlimited set of phrase pairs known from phrase-based machine translation (Koehn et al., 2003). The relation $<_{\mathbf{a}}$ as well as the division into atomic and composed TEUs can straightforwardly be adapted to contiguous translation equivalents.

4 Grammatical translation equivalence

The derivations of a synchronous grammar can be interpreted as deriving a partially ordered set of TEUs as well. A finite derivation $S \rightarrow^{+} \langle \mathbf{s}, \mathbf{t}, \mathbf{a}_G \rangle$ of an instance grammar G is a finite sequence of term-rewritings, where at each step of the sequence a single nonterminal is rewritten using a synchronous production of G. The set of the finite derivations of G defines a language, a set of triples $\langle \mathbf{s}, \mathbf{t}, \mathbf{a}_G \rangle$ consisting of a source string of terminals $\mathbf{s}$, a target string of terminals $\mathbf{t}$ and an alignment between their grammatical constituents. Crucially, the alignment $\mathbf{a}_G$ is obtained by *recursively interpreting* the alignment relations embedded in the synchronous grammar productions in the derivation for all constituents and concerns constituent alignments (as opposed to word alignments).

Grammatical translation equivalents $\mathbf{TE}_G(\mathbf{s}, \mathbf{t})$ A synchronous derivation $S \rightarrow^{+} \langle \mathbf{s}, \mathbf{t}, \mathbf{a}_G \rangle$ can be viewed as a deductive proof that $\langle \mathbf{s}, \mathbf{t}, \mathbf{a}_G \rangle$ is a *grammatical* translation equivalence unit (grammatical TEU). Along the way, a derivation also proves other *constituent-level* (sub-sentential) units as TEUs.

We define a *sub-sentential* grammatical TEU of $\langle \mathbf{s}, \mathbf{t}, \mathbf{a}_G \rangle$ to consist of a triple $\langle \mathbf{s}_x, \mathbf{t}_x, \mathbf{a}_x \rangle$, where $\mathbf{s}_x$ and $\mathbf{t}_x$ are two *subsequences*[3] (of $\mathbf{s}$ and $\mathbf{t}$ respectively), derived synchronously from the same constituent X in some non-empty "tail" of a derivation $S \rightarrow^{+} \langle \mathbf{s}, \mathbf{t}, \mathbf{a}_G \rangle$; importantly, by the workings of G, the alignment $\mathbf{a}_x \subseteq \mathbf{a}_G$ fulfills the requirement that a word in $\mathbf{s}_x$ or in $\mathbf{t}_x$ is linked to another by $\mathbf{a}_G$ iff it is also linked that way by $\mathbf{a}_x$ (i.e., no alignments start out from terminals in $\mathbf{s}_x$ or $\mathbf{t}_x$ and link to terminals outside them). We will denote with $\mathbf{TE}_G(\mathbf{s}, \mathbf{t})$ the *set of all grammatical TEUs* for the sentence pair $\langle \mathbf{s}, \mathbf{t} \rangle$ derived by G.

[3] A subsequence of a string is a subset of the word-position pairs that preserves the order but do not necessarily constitute contiguous substrings.

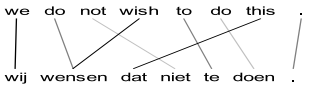

Figure 2: Alignment with both contiguous and discontiguous TEUs (example from Europarl En-Ne).

Subsumption relation $<_{G(\mathbf{s},\mathbf{t})}$ Besides deriving TEUs, a derivation also shows *how* the different TEUs *compose* together into larger TEUs according to the grammar. We are interested in the *subsumption relation*: one grammatical TEU/constituent (u_1) subsumes another (u_2) (written $u_2 <_{G(\mathbf{s},\mathbf{t})} u_1$) iff the latter ($u_2$) is derived within a finite derivation of the former (u_1).[4]

The set of grammatical TEUs for a finite set of derivations for a given sentence pair is the union of the sets defined for the individual derivations. Similarly, the relation between TEU's for a set of derivations is defined as the union of the individual relations.

5 Alignment coverage by intersection

Let a word aligned sentence pair $\langle \mathbf{s}, \mathbf{t}, \mathbf{a} \rangle$ be given, and let us assume that we have a definition of an ordered set $\mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ with partial order $<_{\mathbf{a}}$. We will say that a *grammar formalism covers* $\mathbf{a}$ iff there exists an instance grammar G that fulfills two intersection equations simultaneously:[5]

$$(\mathbf{1}) \quad \mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a}) \cap \mathbf{TE}_G(\mathbf{s}, \mathbf{t}) = \mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$$

$$(\mathbf{2}) \quad <_{\mathbf{a}} \cap <_{G(\mathbf{s},\mathbf{t})} = <_{\mathbf{a}}$$

In the second equation, the intersection of partial orders is based on the standard view that these are in essence also sets of ordered pairs. In practice, it is sufficient to implement an algorithm that shows

[4] Note that we define this relation exhaustively thereby defining the set of paths in synchronous trees derived by the grammar for $\langle \mathbf{s}, \mathbf{t} \rangle$. Hence, the subsumption relation can be seen to define a forest of synchronous trees.

[5] In this work we have restricted this definition to full coverage (i.e., subset) version but it is imaginable that other measures can be based on the cardinality (size) of the intersection in terms of covered TEUs, in following of measures found in (Søgaard and Kuhn, 2009; Søgaard and Wu, 2009). We leave this to future work.

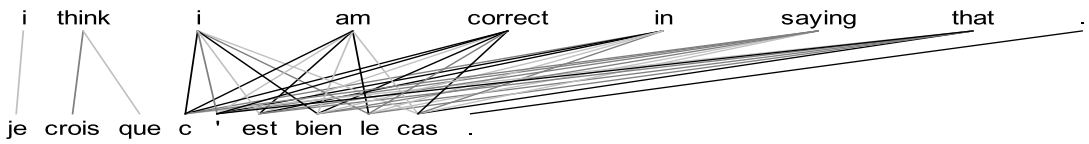

Figure 1: Alignment with only contiguous TEUs (example from LREC En-Fr).

that *G derives every TEU* in $\mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$, and that the subsumption relation $<_{\mathbf{a}}$ between TEUs in $\mathbf{a}$ must be realized by the derivations of G that derive $\mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$. In effect, this way every TEU that subsumes other TEUs must be derived recursively, while the minimal, atomic units (not subsuming any others) must be derived using the lexical productions (endowed with internal word alignments) of NF-ITG. Again, the rationale behind this choice is that the atomic units constitute fixed translation expressions (idiomatic TEUs) which cannot be composed from other TEUs, and hence belong in the lexicon. We will exhibit coverage algorithms for doing so for NF-ITG for the two kinds of semantic interpretations of word alignments.

A note on dedicated instances of NF-ITG Given a translation equivalence definition over word alignments $\mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$, the lexical productions for a *dedicated* instance of NF-ITG are defined[6] by the set $\{X \rightarrow u \mid u \in \mathbf{TE_{Atom}}(\mathbf{s}, \mathbf{t}, \mathbf{a})\}$. This means that the lexical productions have atomic TEUs at the right-hand side including alignments between the words of the source and target terminals. In the sequel, we will only talk about dedicated instances of NF-ITG and hence we will not explicitly repeat this every time.

Given two grammatical TEUs u_1 and u_2, an NF-ITG instance allows their concatenation either in monotone [] or inverted <> order iff they are adjacent on the source and target sides. This fact implies that for every composed translation equivalent $u \in \mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ we can check whether it is derivable by a dedicated NF-ITG instance by checking whether it recursively decomposes into adjacent pairs of TEUs down to the atomic TEUs level. Note that by doing so, we are also implicitly checking whether the subsumption order between the TEUs in $\mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$ is realized by the grammatical derivation (i.e, $<_{G(\mathbf{s},\mathbf{t})} \subseteq <_{\mathbf{a}}$). Formally, an aligned sentence pair $\langle \mathbf{s}, \mathbf{t}, \mathbf{a} \rangle$ is split into a pair of TEUs $\langle \mathbf{s}_1, \mathbf{t}_1, \mathbf{a}_1 \rangle$ and $\langle \mathbf{s}_2, \mathbf{t}_2, \mathbf{a}_2 \rangle$ that can be composed back using the [] and <> productions. If such a split exists, the splitting is conducted recursively for each of $\langle \mathbf{s}_1, \mathbf{t}_1, \mathbf{a}_1 \rangle$ and $\langle \mathbf{s}_2, \mathbf{t}_2, \mathbf{a}_2 \rangle$ until both are atomic TEUs in $\mathbf{TE}(\mathbf{s}, \mathbf{t}, \mathbf{a})$. This recursive splitting is the check of *binarizability* and an algorithm is described in (Huang et al., 2009).

6 A simple algorithm for ITG

We exemplify the grammatical coverage for (normal form) ITG by employing a standard tabular algorithm based on CYK (Younger, 1967). The algorithm works in two phases creating a chart containing TEUs with associated inferences. In the initialization phase (Algorithm 1), for all source spans that correspond to translation equivalents and which have no smaller translation equivalents they contain, *atomic translation equivalents* are added as atomic inferences to the chart. In the second phase, based on the atomic inferences, the simple rules of NF-ITG are applied to add inferences for increasingly larger chart entries. An inference is added (Algorithms 2 and 3) iff a chart entry can be split into two sub-entries for which inferences already exist, and furthermore the union of the sets of target positions for those two entries form a consecutive range.[7] The *addMonotoneInference* and *addInvertedInference* in Algorithm 3 mark the composit inferences by monotone and inverted productions respectively.

[6] Unaligned words add one wrinkle in this scheme: informally, we consider a TEU u formed by attaching unaligned words to an atomic TEU also as atomic iff u is absolutely needed to cover the aligned sentence pair.

[7] We are not treating unaligned words formally here. For unaligned source and target words, we have to generate the different inferences corresponding to different groupings with their neighboring aligned words. Using pre-processing we set aside the unaligned words, then parse the remaining word alignment fully. After parsing, by post-processing, we introduce in the parse table atomic TEUs that include the unaligned words.

```
InitializeChart
Input  : ⟨s, t, a⟩
Output: Initialized chart for atomic units
for spanLength ← 2 to n do
    for i ← 0 to n − spanLength + 1 do
        j ← i + spanLength − 1
        u ← {⟨X, Y⟩ : X ∈ {i...j}}
        if (u ∈ TE_Atom(s, t, a)) then
            addAtomicInference(chart[i][j], u)
        end
    end
end
```

Algorithm 1: Algorithm that initializes the Chart with atomic sub-sentential TEUs. In order to be atomic, a TEU may not contain smaller TEUs that consist of a proper subset of the alignments (and associated words) of the TEU.

```
ComputeTEUsNFITG
Input  : ⟨s, t, a⟩
Output: TRUE/FALSE for coverage
InitializeChart(chart)
for spanLength ← 2 to n do
    for i ← 0 to n − spanLength + 1 do
        j ← i + spanLength − 1
        if chart[i][j] ∈ TE(s, t, a) then
            continue
        end
        for splitPoint ← i + 1 to j do
            a′ ← (chart[i][k − 1] ∪ chart[k][j])
            if (chart[i][k − 1] ∈ TE(s, t, a)) ∧
            (chart[k][j] ∈ TE(s, t, a)) ∧
            (a′ ∈ TE(s, t, a)) then
                addTEU(chart, i, j, k, a′)
            end
        end
    end
    if (chart[0][n − 1] ≠ ∅) then
        return TRUE
    else
        return FALSE
    end
end
```

Algorithm 2: Algorithm that incrementally builds composite TEUs using only the rules allowed by NF-ITG

```
addTEU
Input  :
 chart - the chart
 i,j,k - the lower, upper and split point indices
 a′ - the TEU to be added
Output: chart with TEU a′ added in the
        intended entry
if Max_{Y_t}({Y_t : ⟨X_s, Y_t⟩ ∈ chart[i][k − 1]})
< Max_{Y_t}({Y_t : ⟨X_s, Y_t⟩ ∈ chart[k][j]}) then
    addMonotoneInference(chart[i][j], a′)
else
    addInvertedInference(chart[i][j], a′)
end
```

Algorithm 3: Algorithm that adds a TEU and associated Inference to the chart

7 Experiments

Data Sets We use manually and automatically aligned corpora. Manually aligned corpora come from two datasets. The first (Graça et al., 2008) consists of six language pairs: Portuguese–English, Portuguese–French, Portuguese–Spanish, English–Spanish, English–French and French–Spanish. These datasets contain 100 sentence pairs each and distinguish *Sure* and *Possible* alignments. Following (Søgaard and Kuhn, 2009), we treat these two equally. The second manually aligned dataset (Padó and Lapata, 2006) contains 987 sentence pairs from the English-German part of Europarl annotated using the Blinker guidelines (Melamed, 1998). The automatically aligned data comes from Europarl (Koehn, 2005) in three language pairs (English–Dutch, English–French and English–German). The corpora are automatically aligned using GIZA++ (Och and Ney, 2003) in combination with the grow-diag-final-and heuristic. With sentence length cut-off 40 on both sides these contain respectively 945k, 949k and 995k sentence pairs.

Grammatical Coverage (GC) is defined as the percentage word alignments (sentence pairs) in a parallel corpus that can be covered by an instance of the grammar (NF-ITG) (cf. Section 5). Clearly, GC depends on the chosen semantic interpretation of word alignments: contiguous TE's (phrase pairs) or discontiguous TE's.

Alignments Set	GC contiguous TEs	GC discontiguous TEs
Hand aligned corpora		
English–French	76.0	75.0
English–Portuguese	78.0	78.0
English–Spanish	83.0	83.0
Portuguese–French	78.0	74.0
Portuguese–Spanish	91.0	91.0
Spanish–French	79.0	74.0
LREC Corpora Average	80.83±5.49	79.17±6.74
English–German	45.427	45.325
Automatically aligned Corpora		
English–Dutch	45.533	43.57
English–French	52.84	49.95
English–German	45.59	43.72
Automatically aligned corpora average	47.99±4.20	45.75±3.64

Table 1: The grammatical coverage (GC) of NF-ITG for different corpora dependent on the interpretation of word alignments: contiguous Translation Equivalence or discontiguous Translation Equivalence

Results Table 1 shows the Grammatical Coverage (GC) of NF-ITG for the different corpora dependent on the two alternative definitions of *translation equivalence*. The first thing to notice is that there is just a small difference between the Grammatical Coverage scores for these two definitions. The difference is in the order of a few percentage points, the largest difference is seen for Portuguese–French (79% v.s 74% Grammatical Coverage), for some language pairs there is no difference. For the automatically aligned corpora the absolute difference is on average about 2%. We attribute this to the fact that there are only very few discontiguous TEUs that can be covered by NF-ITG in this data.

The second thing to notice is that the scores are much higher for the corpora from the LREC dataset than they are for the manually aligned English–German corpus. The approximately double source and target length of the manually aligned English–German corpus, in combination with somewhat less dense alignments makes this corpus much harder than the LREC corpora. Intuitively, one would expect that more alignment links make alignments more complicated. This turns out to not always be the case. Further inspection of the LREC alignments also shows that these alignments often consist of parts that are *completely linked*. Such completely linked parts are by definition treated as atomic TEUs, which could make the alignments look simpler. This contrasts with the situation in the manually aligned English–German corpus where on average less alignment links exist per word. Examples 1 and 2 show that dense alignments can be simpler than less dense ones. This is because sometimes the density implies idiomatic TEUs which leads to rather flat lexical productions. We think that idiomatic TEUs reasonably belong in the lexicon.

When we look at the results for the automatically aligned corpora at the lowest rows in the table, we see that these are comparable to the results for the manually aligned English–German corpus (and much lower than the results for the LREC corpora). This could be explained by the fact that the manually aligned English–German is not only Europarl data, but possibly also because the manual alignments themselves were obtained by initialization with the GIZA++ alignments. In any case, the manually and automatically acquired alignments for this data are not too different from the perspective of NF-ITG. Further differences might exist if we would employ another class of grammars, e.g., full SCFGs.

One the one hand, we find that manual alignments are well but not fully covered by NF-ITG. On the other, the automatic alignments are not covered well but NF-ITG. This suggests that these automatic alignments are difficult to cover by NF-ITG, and the reason could be that these alignments are built heuristically by trading precision for recall cf.

(Och and Ney, 2003). Sogaard (Søgaard, 2010) reports that full ITG provides a few percentage points gains over NF-ITG.

Overall, we find that our results for the LREC data are far higher Sogaard's (Søgaard, 2010) results but lower than the upperbounds of (Søgaard and Wu, 2009). A similar observation holds for the English–German manually aligned EuroParl data, albeit the maximum length (15) used in (Søgaard and Wu, 2009; Søgaard, 2010) is different from ours (40). We attribute the difference between our results and Sogaard's approach to our choice to adopt lexical productions of NF-ITG that contain own internal alignments (the detailed version) and determined by the atomic TEUs of the word alignment. Our results differ substantially from (Søgaard and Wu, 2009) who report upperbounds (indeed our results still fall within these upperbounds for the LREC data).

8 Related Work

The array of work described in (Zens and Ney, 2003; Wellington et al., 2006; Søgaard and Wu, 2009; Søgaard and Kuhn, 2009; Søgaard, 2010) concentrates on methods for calculating *upperbounds* on the alignment coverage for all ITGs, including NF-ITG. Interestingly, these upperbounds are determined by *filtering/excluding complex alignment phenomena* known formally to be beyond (NF-)ITG. None of these earlier efforts discussed explicitly the dilemmas of instantiating a grammar formalism or how to formally parse word alignments.

The work in (Zens and Ney, 2003; Søgaard and Wu, 2009), defining and counting TEUs, provides a far tighter upperbound than (Wellington et al., 2006), who use the disjunctive interpretation of word alignments, interpreting multiple alignment links of the same word as alternatives. We adopt the conjunctive interpretation of word alignments like a majority of work in MT, e.g., (Ayan and Dorr, 2006; Fox, 2002; Søgaard and Wu, 2009; Søgaard, 2010).

In deviation from earlier work, the work in (Søgaard and Kuhn, 2009; Søgaard and Wu, 2009; Søgaard, 2010) discusses TEUs defined over word alignments explicitly, and defines evaluation metrics based on TEUs. In particular, Sogaard (Søgaard, 2010) writes that he employs "a more aggressive search" for TEUs than earlier work, thereby leading to far tighter upperbounds on hand aligned data. Our results seem to back this claim but, unfortunately, we could not pin down the formal details of his procedure.

More remotely related, the work described in (Huang et al., 2009) presents a binarization algorithm for productions of an SCFG instance (as opposed to formalism). Although somewhat related, this is different from checking whether there exists an NF-ITG instance (which has to be determined) that covers a word alignment.

In contrast with earlier work, we present the alignment coverage problem as an intersection of two partially ordered sets (graphs). The partial order over TEUs as well as the formal definition of parsing as intersection in this work are novel elements, making explicit the view of word alignments as automata generating partially order sets.

9 Conclusions

In this paper we provide a formal characterization for the problem of determining the coverage of a word alignment by a given grammar formalism as the intersection of two partially ordered sets. These partially ordered set of TEUs can be formalized in terms of hyper-graphs implementing forests (packed synchronous trees), and the coverage as the intersection between sets of synchronous trees generalizing the trees of (Zhang et al., 2008).

Practical explorations of our findings for the benefit of models of learning reordering are underway. In future work we would like to investigate the extension of this work to other limited subsets of SCFGs. We will also investigate the possibility of devising ITGs with explicit links between terminal symbols in the productions, exploring different kinds of linking.

Acknowledgements We thank reviewers for their helpful comments, and thank Mark-Jan Nederhof for illuminating discussions on parsing as intersection. This work is supported by The Netherlands Organization for Scientific Research (NWO) under grant nr. 612.066.929.

References

Nacip Ayan and Bonnie Dorr. 2006. Going beyond AER: an extensive analysis of word alignments and their impact on MT. In *Proc. of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pages 9–16, Morristown, NJ, USA.

Yehoshua Bar-Hillel, Micha Perles, and Eli Shamir. 1964. On formal properties of simple phrase structure grammars. In Y. Bar-Hillel, editor, *Language and Information: Selected Essays on their Theory and Application*, chapter 9, pages 116–150. Addison-Wesley, Reading, Massachusetts.

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A gibbs sampler for phrasal synchronous grammar induction. In *ACL/AFNLP*, pages 782–790.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, page 61–72.

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

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

Trevor Cohn and Phil Blunsom. 2009. A bayesian model of syntax-directed tree to string grammar induction. In *EMNLP*, pages 352–361.

John DeNero, Daniel Gillick, James Zhang, and Dan Klein. 2006. Why generative phrase models underperform surface heuristics. In *Proceedings of the workshop on SMT*, pages 31–38.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, Proceedings of EMNLP, pages 304–311, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joao Graça, Joana Pardal, Luísa Coheur, and Diamantino Caseiro. 2008. Building a golden collection of parallel multi-language word alignment. In *LREC'08*, Marrakech, Morocco. European Language Resources Association (ELRA).

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised itg models. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 923–931, Suntec, Singapore, August. Association for Computational Linguistics.

Liang Huang, Hao Zhang, Daniel Gildea, and Kevin Knight. 2009. Binarization of synchronous context-free grammars. *Computational Linguistics*, 35(4):559–595.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Human Language Technology Conference, HLT-NAACL*, May.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT Summit*.

Bernard Lang. 1988. Parsing incomplete sentences. In *Proceedings of COLING*, pages 365–371.

J. Scott McCarley, Abraham Ittycheriah, Salim Roukos, Bing Xiang, and Jian-Ming Xu. 2011. A correction model for word alignments. In *Proceedings of EMNLP*, pages 889–898.

Dan Melamed. 1998. Annotation style guide for the blinker project, version 1.0. Technical Report IRCS TR #98-06, University of Pennsylvania.

Markos Mylonakis and Khalil Sima'an. 2011. Learning hierarchical translation structure with linguistic annotations. In *Proceedings of the HLT/NAACL-2011*.

Mark-Jan Nederhof and Giorgio Satta. 2004. The language intersection problem for non-recursive context-free grammars. *Inf. Comput.*, 192(2):172–184.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *ACL-COLING'06*, ACL-44, pages 1161–1168, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jason Riesa and Daniel Marcu. 2010. Hierarchical search for word alignment. In *Proceedings of ACL*, pages 157–166.

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 i n Natural Language Processing*, pages 803–810, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Anders Søgaard and Jonas Kuhn. 2009. Empirical lower bounds on alignment error rates in syntax-based machine translation. In *SSST '09*, pages 19–27, Stroudsburg, PA, USA. 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 Workshop on Parsing Technologies (IWPT-2009), 7-9 October 2009, Paris, France*,

pages 33–36. The Association for Computational Linguistics.

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)*.

Roy Tromble and Jason Eisner. 2009. Learning linear ordering problems for better translation. In *Proceedings of EMNLP'09*, pages 1007–1016, Singapore.

Gertjan van Noord. 1995. The intersection of finite state automata and definite clause grammars. In *Proceedings of ACL*, pages 159–165.

Benjamin Wellington, Sonjia Waxmonsky, and I. Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

D.H. Younger. 1967. Recognition and parsing of context-free languages in time n^3. *Information and Control*, 10(2):189–208.

Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the Annual Meeting of the ACL*, pages 144–151.

Hao Zhang, Daniel Gildea, and David Chiang. 2008. Extracting synchronous grammar rules from word-level alignments in linear time. In *Proceedings of COLING*, pages 1081–1088.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the North-American Chapter of the ACL (NAACL'06)*, pages 138–141.