The Hidden TAG Model: Synchronous Grammars for Parsing Resource-Poor Languages

David Chiang*
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292, USA
chiang@isi.edu

Owen Rambow
Center for Computational Learning Systems
Columbia University
475 Riverside Dr., Suite 850
New York, NY, USA
rambow@cs.columbia.edu

Abstract

This paper discusses a novel probabilistic synchronous TAG formalism, synchronous Tree Substitution Grammar with sister adjunction (TSG+SA). We use it to parse a language for which there is no training data, by leveraging off a second, related language for which there is abundant training data. The grammar for the resource-rich side is automatically extracted from a treebank; the grammar on the resource-poor side and the synchronization are created by handwritten rules. Our approach thus represents a combination of grammar-based and empirical natural language processing. We discuss the approach using the example of Levantine Arabic and Standard Arabic.

1 Parsing Arabic Dialects and Tree Adjoining Grammar

The Arabic language is a collection of spoken dialects and a standard written language. The standard written language is the same throughout the Arab world, Modern Standard Arabic (MSA), which is also used in some scripted spoken communication (news casts, parliamentary debates). It is based on Classical Arabic and is not a native language of any Arabic speaking people, i.e., children do not learn it from their parents but in school. Thus most native speakers of Arabic are unable to produce sustained spontaneous MSA. The dialects show phonological, morphological, lexical, and syntactic differences comparable to those among the Romance languages. They vary not only along a geographical continuum but also with other sociolinguistic variables such as the urban/rural/Bedouin dimension.

The multidialectal situation has important negative consequences for Arabic natural language processing (NLP): since the spoken dialects are not officially written and do not have standard orthography, it is very costly to obtain adequate corpora, even unannotated corpora, to use for training NLP tools such as parsers. Furthermore, there are almost no parallel corpora involving one dialect and MSA.

The question thus arises how to create a statistical parser for an Arabic dialect, when statistical parsers are typically trained on large corpora of parse trees. We present one solution to this problem, based on the assumption that it is easier to manually create new resources that relate a dialect to MSA (lexicon and grammar) than it is to manually create syntactically annotated corpora in the dialect. In this paper, we deal with Levantine Arabic (LA). Our approach does not assume the existence of any annotated LA corpus (except for development and testing), nor of a parallel LA-MSA corpus.

The approach described in this paper uses a special parameterization of stochastic synchronous TAG (Shieber, 1994) which we call a "hidden TAG model." This model couples a model of MSA trees, learned from the Arabic Treebank, with a model of MSA-LA translation, which is initialized by hand and then trained in an unsupervised fashion. Parsing new LA sentences then entails simultaneously building a forest of MSA trees and the corresponding forest of LA trees. Our implementation uses an extension of our monolingual parser (Chiang, 2000) based on tree-substitution

*This work was primarily carried out while the first author was at the University of Maryland Institute for Advanced Computer Studies.

Proceedings of the 8th International Workshop on Tree Adjoining Grammar and Related Formalisms, pages 1–8,
Sydney, July 2006.

grammar with sister adjunction (TSG+SA).

The main contributions of this paper are as follows:

1. We introduce the novel concept of a hidden TAG model.
2. We use this model to combine statistical approaches with grammar engineering (specifically motivated from the linguistic facts). Our approach thus exemplifies the specific strength of a grammar-based approach.
3. We present an implementation of stochastic synchronous TAG that incorporates various facilities useful for training on real-world data: sister-adjunction (needed for generating the flat structures found in most treebanks), smoothing, and Inside-Outside reestimation.

This paper is structured as follows. We first briefly discuss related work (Section 2) and some of the linguistic facts that motivate this work (Section 3). We then present the formalism, probabilistic model, and parsing algorithm (Section 4). Finally, we discuss the manual grammar engineering (Section 5) and evaluation (Section 6).

2 Related Work

This paper is part of a larger investigation into parsing Arabic dialects (Rambow et al., 2005; Chiang et al., 2006). In that investigation, we examined three different approaches:

- Sentence transduction, in which a dialect sentence is roughly translated into one or more MSA sentences and then parsed by an MSA parser.
- Treebank transduction, in which the MSA treebank is transduced into an approximation of a LA treebank, on which a LA parer is then trained.
- Grammar transduction, which is the name given in the overview papers to the approach discussed in this paper. The present paper provides for the first time a complete technical presentation of this approach.

Overall, grammar transduction outperformed the other two approaches.

In other work, there has been a fair amount of interest in parsing one language using another language, see for example (Smith and Smith, 2004; Hwa et al., 2004). Much of this work, like ours, relies on synchronous grammars (CFGs). However, these approaches rely on parallel corpora. For MSA and its dialects, there are no naturally occurring parallel corpora. It is this fact that has led us to investigate the use of explicit linguistic knowledge to complement machine learning.

3 Linguistic Facts

We illustrate the differences between LA and MSA using an example:

(1) a. الرجال بيحبو ش الشغل هدا (LA)

AlrjAl byHbw $ Al$gl hdA
the-men like not the-work this

the men do not like this work

b. لا يحب الرجال هذا العمل (MSA)

lA yHb AlrjAl h*A AlEml
not like the-men this the-work

the men do not like this work

Lexically, we observe that the word for 'work' is الشغل *Al$gl* in LA but العمل *AlEml* in MSA. In contrast, the word for 'men' is the same in both LA and MSA: الرجال *AlrjAl*. There are typically also differences in function words, in our example ش *$* (LA) and لا *lA* (MSA) for 'not'. Morphologically, we see that LA بيحبو *byHbw* has the same stem as MA يحب *yHb*, but with two additional morphemes: the present aspect marker *b-* which does not exist in MSA, and the agreement marker *-w*, which is used in MSA only in subject-initial sentences, while in LA it is always used.

Syntactically, we observe three differences. First, the subject precedes the verb in LA (SVO order), but follows in MSA (VSO order). This is in fact not a strict requirement, but a strong preference: both varieties allow both orders, but in the dialects, the SVO order is more common, while in MSA, the VSO order is more common. Second, we see that the demonstrative determiner follows the noun in LA, but precedes it in MSA. Finally, we see that the negation marker follows the verb in LA, while it precedes the verb in MSA. (Levantine also has other negation markers that precede the verb, as well as the circumfix *m- -$*.) The two phrase structure trees are shown in Figure 1 in the convention of the Linguistic Data Consortium (Maamouri et al., 2004). Unlike the phrase

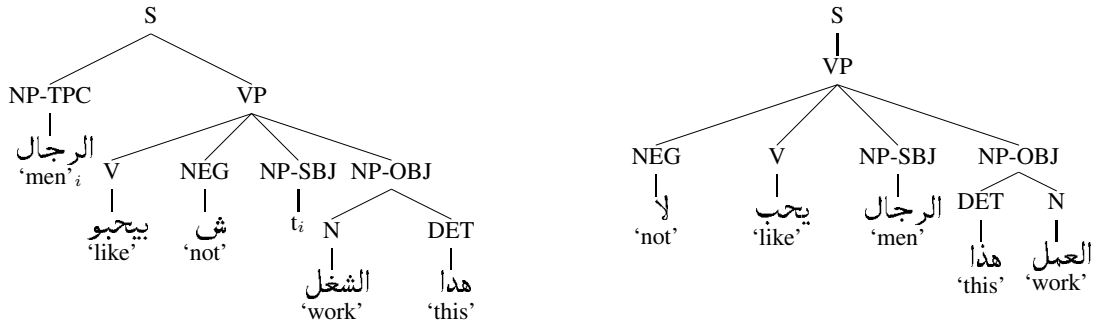

Figure 1: LDC-style left-to-right phrase structure trees for LA (left) and MSA (right) for sentence (1)

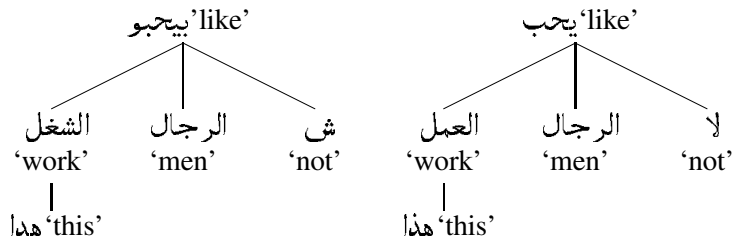

Figure 2: Unordered dependency trees for LA (left) and MSA (right) for sentence (1)

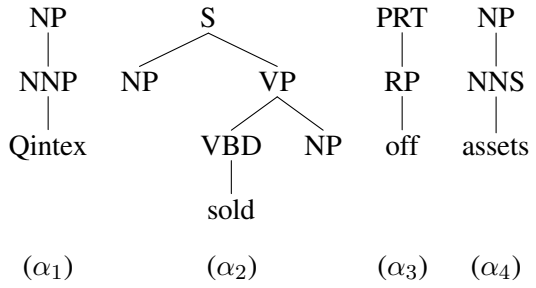

Figure 3: Example elementary trees.

structure trees, the (unordered) dependency trees for the MSA and LA sentences are isomorphic, as shown in Figure 2. They differ only in the node labels.

4 Model

4.1 The synchronous TSG+SA formalism

Our parser (Chiang, 2000) is based on synchronous tree-substitution grammar with sister-adjunction (TSG+SA). Tree-substitution grammar (Schabes, 1990) is TAG without auxiliary trees or adjunction; instead we include a weaker composition operation, *sister-adjunction* (Rambow et al., 2001), in which an initial tree is inserted between two sister nodes (see Figure 4). We allow multiple sister-adjunctions at the same site, similar to how Schabes and Shieber (1994) allow multiple adjunctions of modifier auxiliary trees.

A *synchronous* TSG+SA is a set of pairs of elementary trees. In each pair, there is a one-to-one correspondence between the substitution/sister-adjunction sites of the two trees, which we represent using boxed indices (Figure 5). A derivation then starts with a pair of initial trees and proceeds by substituting or sister-adjoining elementary tree pairs at coindexed sites. In this way a set of string pairs $\langle S, S' \rangle$ is generated.

Sister-adjunction presents a special problem for synchronization: if multiple tree pairs sister-adjoin at the same site, how should their order on the source side relate to the order on the target side? Shieber's solution (Shieber, 1994) is to allow any ordering. We adopt a stricter solution: for each pair of sites, fix a permutation (either identity or reversal) for the tree pairs that sister-adjoin there. Owing to the way we extract trees from the Treebank, the simplest choice of permutations is: if the two sites are both to the left of the anchor or both to the right of the anchor, then multiple sister-adjoined tree pairs will appear in the same order on both sides; otherwise, they will appear in the opposite order. In other words, multiple sister-adjunction always adds trees from the anchor outward.

A *stochastic* synchronous TSG+SA adds probabilities to the substitution and sister-adjunction operations: the probability of substituting an elementary tree pair $\langle \alpha, \alpha' \rangle$ at a substitution site pair

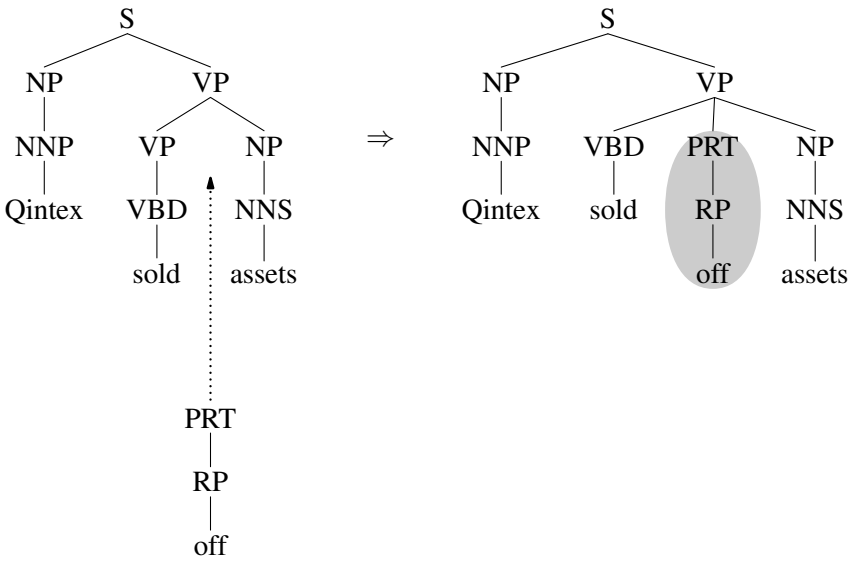

Figure 4: Sister-adjunction, with inserted material shown with shaded background

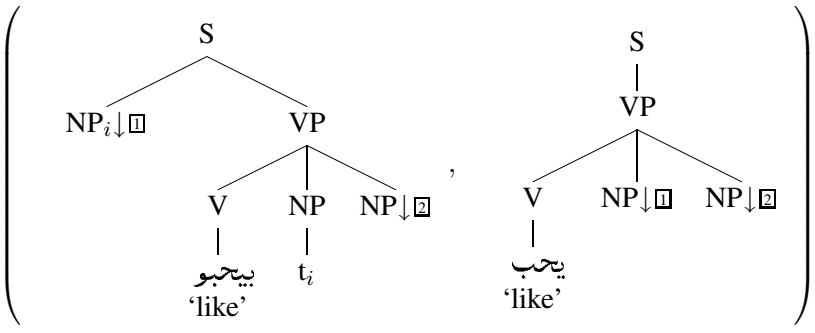

Figure 5: Example elementary tree pair of a synchronous TSG: the **SVO** transformation (LA on left, MSA on right)

$\langle \eta, \eta' \rangle$ is $P_{\mathbf{s}}(\alpha, \alpha' \mid \eta, \eta')$, and the probability of sister-adjoining $\langle \alpha, \alpha' \rangle$ at a sister-adjunction site pair $\langle \eta, i, \eta', i' \rangle$ is $P_{\mathbf{sa}}(\alpha, \alpha' \mid \eta, i, \eta', i')$, where i and i' indicate that the sister-adjunction occurs between the i and $(i+1)$st (or i' and $(i'+1)$st) sisters. These parameters must satisfy the normalization conditions

$$\sum_{\alpha,\alpha'} P_{\mathbf{s}}(\alpha, \alpha' \mid \eta, \eta') = 1 \quad (1)$$

$$\sum_{\alpha,\alpha'} P_{\mathbf{sa}}(\alpha, \alpha' \mid \eta, i, \eta', i') + P_{\mathbf{sa}}(\textbf{STOP} \mid \eta, i, \eta', i') = 1 \quad (2)$$

4.2 Parsing by translation

We intend to apply a stochastic synchronous TSG+SA to input sentences S'. This requires projecting any constraints from the unprimed side of the synchronous grammar over to the primed side, and then parsing the sentences S' using the projected grammar, using a straightforward generalization of the CKY and Viterbi algorithms. This gives the highest-probability derivation of the synchronous grammar that generates S' on the primed side, which includes a parse for S' and, as a by-product, a parsed translation of S'.

Suppose that S' is a sentence of LA. For the present task we are not actually interested in the MSA translation of S', or the parse of the MSA translation; we are only interested in the parse of S'. The purpose of the MSA side of the grammar is to provide reliable statistics. Thus, we approximate the synchronous rewriting probabilities as:

$$P_{\mathbf{s}}(\alpha, \alpha' \mid \eta, \eta') \approx P_s(\alpha \mid \eta) P_t(\alpha' \mid \alpha) \quad (3)$$

$$P_{\mathbf{sa}}(\alpha, \alpha' \mid \eta, i, \eta', i') \approx P_{sa}(\alpha \mid \eta, i) P_t(\alpha' \mid \alpha) \quad (4)$$

These factors, as we will see shortly, are much easier to estimate given the available resources.

This factorization is analogous to a hidden Markov model: the primed derivation is the observation, the unprimed derivation is the hidden state sequence (except it is a branching process instead of a chain); the P_s and P_{sa} are like the transition probabilities and the P_t are like the observation probabilities. Hence, we call this model a "hidden TAG model."

4.3 Parameter estimation and smoothing

P_s and P_{sa} are the parameters of a monolingual TSG+SA and can be learned from a monolingual Treebank (Chiang, 2000); the details are not important here.

As for P_t, in order to obtain better probability estimates, we further decompose P_t into P_{t1} and P_{t2} so they can be estimated separately (as in the monolingual parsing model):

$$P_t(\alpha' \mid \alpha) \approx P_{t1}(\bar{\alpha}' \mid \bar{\alpha}, w', t', w, t) \times P_{t2}(w', t' \mid w, t) \quad (5)$$

where w and t are the lexical anchor of α and its POS tag, and $\bar{\alpha}$ is the equivalence class of α modulo lexical anchors and their POS tags. P_{t2} represents the lexical transfer model, and P_{t1} the syntactic transfer model. P_{t1} and P_{t2} are initially assigned by hand; P_{t1} is then reestimated by EM.

Because the full probability table for P_{t1} would be too large to write by hand, and because our training data might be too sparse to reestimate it well, we smooth it by approximating it as a linear combination of backoff models:

$$\begin{aligned} P_{t1}(\bar{\alpha}' \mid \bar{\alpha}, w', t', w, t) \approx \\ \lambda_1 P_{t11}(\bar{\alpha}' \mid \bar{\alpha}, w', t', w, t) + \\ (1-\lambda_1)(\lambda_2 P_{t12}(\bar{\alpha}' \mid \bar{\alpha}, w', t') + \\ (1-\lambda_2) P_{t13}(\bar{\alpha}' \mid \bar{\alpha})) \end{aligned} \quad (6)$$

where each λ_i, unlike in the monolingual parser, is simply set to 1 if an estimate is available for that level, so that it completely overrides the further backed-off models.

The initial estimates for the P_{t1i} are set by hand. The availability of three backoff models makes it easy to specify the initial guesses at an appropriate level of detail: for example, one might give a general probability of some $\bar{\alpha}$ mapping to $\bar{\alpha}'$ using P_{t13}, but then make special exceptions for particular lexical anchors using P_{t11} or P_{t12}.

Finally P_{t2} is reestimated by EM on some held-out unannotated sentences of L', using the same method as Chiang and Bikel (2002) but on the syntactic transfer probabilities instead of the monolingual parsing model. Another difference is that, following Bikel (2004), we do not recalculate the λ_i at each iteration, but use the initial values throughout.

5 A Synchronous TSG-SA for Dialectal Arabic

Just as the probability model discussed in the preceding section factored the rewriting probabilities

into three parts, we create a synchronous TSG-SA and the probabilities of a hidden TAG model in three steps:

- P_s and P_{sa} are the parameters of a monolingual TSG+SA for MSA. We extract a grammar for the resource-rich language (MSA) from the Penn Arabic Treebank in a process described by Chiang and others (Chiang, 2000; Xia et al., 2000; Chen, 2001).
- For the lexical transfer model P_{t2}, we create by hand a probabilistic mapping between $(\text{word}, \text{POS tag})$ pairs in the two languages.
- For the syntactic transfer model P_{t1}, we created by hand a grammar for the resource-poor language and a mapping between elementary trees in the two grammars, along with initial guesses for the mapping probabilities.

We discuss the hand-crafted lexicon and synchronous grammar in the following subsections.

5.1 Lexical Mapping

We used a small, hand-crafted lexicon of 100 words which mapped all LA function words and some of the most common open-class words to MSA. We assigned uniform probabilities to the mapping. All other MSA words were assumed to also be LA words. Unknown LA words were handled using the standard unknown word mechanism.

5.2 Syntactic Mapping

Because of the underlying syntactic similarity between the two varieties of Arabic, we assume that every tree in the MSA grammar extracted from the MSA treebank is also a LA tree. In addition, we define tree transformations in the Tsurgeon package (Levy and Andrew, 2006). These consist of a pattern which matches MSA elementary trees in the extracted grammar, and a transformation which produces a LA elementary tree. We perform the following tree transformations on all elementary trees which match the underlying MSA pattern. Thus, each MSA tree corresponds to at least two LA trees: the original one and the transformed one. If several transformations apply, we obtain multiple transformed trees.

- Negation (**NEG**): we insert a *$* negation marker immediately following each verb. The preverbal marker is generated by a lexical translation of an MSA elementary tree.
- VSO-SVO Ordering (**SVO**): Both Verb-Subject-Object (VSO) and Subject-Verb-Object (SVO) constructions occur in MSA and LA treebanks. But pure VSO constructions (without pro-drop) occur in the LA corpus only 10ordering in MSA. Hence, the goal is to skew the distributions of the SVO constructions in the MSA data. Therefore, VSO constructions are replicated and converted to SVO constructions. One possible resulting pair of trees is shown in Figure 5.
- The *bd* construction (**BD**): *bd* is a LA noun that means 'want'. It acts like a verb in verbal constructions yielding VP constructions headed by NN. It is typically followed by an enclitic possessive pronoun. Accordingly, we defined a transformation that translated all the verbs meaning 'want'/'need' into the noun *bd* and changed their respective POS tag to NN. The subject clitic is transformed into a possessive pronoun clitic. Note that this construction is a combination lexical and syntactic transformation, and thus specifically exploits the extended domain of locality of TAG-like formalisms. One possible resulting pair of trees is shown in Figure 6.

6 Experimental Results

While our approach does not rely on any annotated corpus for LA, nor on a parallel corpus MSA-LA, we use a small treebank of LA (Maamouri et al., 2006) to analyze and test our approach. The LA treebank is divided into a development corpus and a test corpus, each about 11,000 tokens (using the same tokenization scheme as employed in the MSA treebank).

We first use the development corpus to determine which of the transformations are useful. We use two conditions. In the first, the input text is not tagged, and the parser hypothesizes tags. In the second, the input text is tagged with the gold (correct) tag. The results are shown in Table 1. The baseline is simply the application of a pure MSA Chiang parser to LA. We see that important improvements are obtained using the lexical mapping. Adding the **SVO** transformation does not improve the results, but the **NEG** and **BD** transformations help slightly, and their effect is (partly)

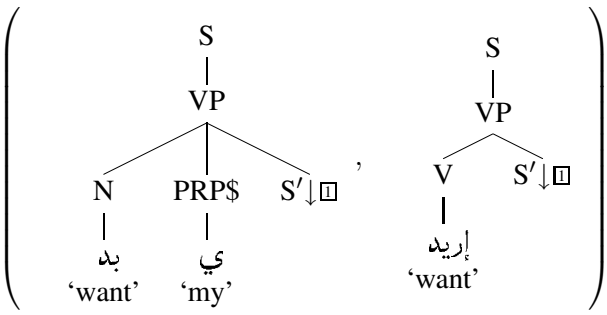

Figure 6: Example elementary tree pair of a synchronous TSG: the **BD** transformation (LA on left, MSA on right)

cumulative. (We did not perform these tuning experiments on input without POS tags.)

The evaluation on the test corpus confirms these results. Using the **NEG** and **BD** transformations and the small lexicon, we obtain a 17.3% error reduction relative to the baseline parser (Figure 2).

These results show that the translation lexicon can be integrated effectively into our synchronous grammar framework. In addition, some syntactic transformations are useful. The **SVO** transformation, we assume, turns out not to be useful because the **SVO** word order is also possible in MSA, so that the new trees were not needed and needlessly introduced new derivations. The **BD** transformation shows the importance not of general syntactic transformations, but rather of lexically specific syntactic transformations: varieties within one language family may differ more in terms of the lexico-syntactic constructions used for a specific (semantic or pragmatic) purpose than in their basic syntactic inventory. Note that our tree-based synchronous formalism is ideally suited for expressing such transformations since it is lexicalized, and has an extended domain of locality. Given the impact of the **BD** transformation, in future work we intend to determine more lexico-structural transformations, rather than pure syntactic transformations. However, one major impediment to obtaining better results is the disparity in genre and domain which affects the overall performance.

7 Conclusion

We have presented a new probabilistic synchronous TAG formalism, synchronous Tree Substitution Grammar with sister adjunction (TSG+SA). We have introduced the concept of a hidden TAG model, analogous to a Hidden Markov Model. It allows us to parse a resource-poor language using a treebank-extracted probabilistic grammar for a resource-rich language, along with a hand-crafted synchronous grammar for the resource-poor language. Thus, our model combines statistical approaches with grammar engineering (specifically motivated from the linguistic facts). Our approach thus exemplifies the specific strength of a grammar-based approach. While we have applied this approach to two closely related languages, it would be interesting to apply this approach to more distantly related languages in the future.

Acknowledgments

This paper is based on work done at the 2005 Johns Hopkins Summer Workshop, which was partially supported by the National Science Foundation under grant 0121285. The first author was additionally supported by ONR MURI contract FCPO.810548265 and Department of Defense contract RD-02-5700. The second author was additionally supported by contract HR0011-06-C-0023 under the GALE program. We wish to thank the other members of our JHU team (our co-authors on (Rambow et al., 2005)), especially Nizar Habash and Mona Diab for their help with the Arabic examples, and audiences at JHU for their useful feedback.

References

Daniel M. Bikel. 2004. *On the Parameter Space of Generative Lexicalized Parsing Models*. Ph.D. thesis, University of Pennsylvania.

John Chen. 2001. *Towards Efficient Statistical Parsing Using Lexicalized Grammatical Information*. Ph.D. thesis, University of Delaware.

	no tags			gold tags		
	LP	LR	F1	LP	LR	F1
Baseline	59.4	51.9	55.4	64.0	58.3	61.0
Lexical	63.0	60.8	61.9	66.9	67.0	66.9
+ SVO				66.9	66.7	66.8
+ NEG				67.0	67.0	67.0
+ BD				67.4	67.0	67.2
+ NEG + BD				67.4	67.1	67.3

Table 1: Results on development corpus: LP = labeled precision, LR = labeled recall, F1 = balanced F-measure

	no tags	gold tags
	F1	F1
Baseline	53.5	60.2
Lexical + NEG + BD	60.2	67.1

Table 2: Results on the test corpus: F1 = balanced F-measure

David Chiang and Daniel M. Bikel. 2002. Recovering latent information in treebanks. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING)*, pages 183–189.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *Proceedings of EACL*.

David Chiang. 2000. Statistical parsing with an automatically-extracted tree adjoining grammar. In *38th Meeting of the Association for Computational Linguistics (ACL'00)*, pages 456–463, Hong Kong, China.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2004. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of LREC*.

Mohamed Maamouri, Ann Bies, and Tim Buckwalter. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, and Dalila Tabessi. 2006. Developing and using a pilot dialectal Arabic treebank. In *Proceedings of LREC*, Genoa, Italy.

Owen Rambow, K. Vijay-Shanker, and David Weir. 2001. D-Tree Substitution Grammars. *Computational Linguistics*, 27(1).

Owen Rambow, David Chiang, Mona Diab, Nizar Habash, Rebecca Hwa, Khalil Sima'an, Vincent Lacey, Roger Levy, Carol Nichols, and Safiullah Shareef. 2005. Parsing Arabic dialects. Final Report, 2005 JHU Summer Workshop.

Yves Schabes and Stuart Shieber. 1994. An alternative conception of tree-adjoining derivation. *Computational Linguistics*, 1(20):91–124.

Yves Schabes. 1990. *Mathematical and Computational Aspects of Lexicalized Grammars.* Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

Stuart B. Shieber. 1994. Restricting the weak generative capacity of Synchronous Tree Adjoining Grammar. *Computational Intelligence*, 10(4):371–385.

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP04)*.

Fei Xia, Martha Palmer, and Aravind Joshi. 2000. A uniform method of grammar extraction and its applications. In *Proceedings of the 2000 Conference on Empirical Methods in Natural Language Processing (EMNLP00)*, Hong Kong.