Parsing Models for Identifying Multiword Expressions

Spence Green*
Stanford University

Marie-Catherine de Marneffe**
Stanford University

Christopher D. Manning†
Stanford University

Multiword expressions lie at the syntax/semantics interface and have motivated alternative theories of syntax like Construction Grammar. Until now, however, syntactic analysis and multiword expression identification have been modeled separately in natural language processing. We develop two structured prediction models for joint parsing and multiword expression identification. The first is based on context-free grammars and the second uses tree substitution grammars, a formalism that can store larger syntactic fragments. Our experiments show that both models can identify multiword expressions with much higher accuracy than a state-of-the-art system based on word co-occurrence statistics.

We experiment with Arabic and French, which both have pervasive multiword expressions. Relative to English, they also have richer morphology, which induces lexical sparsity in finite corpora. To combat this sparsity, we develop a simple factored lexical representation for the context-free parsing model. Morphological analyses are automatically transformed into rich feature tags that are scored jointly with lexical items. This technique, which we call a factored lexicon, improves both standard parsing and multiword expression identification accuracy.

1. Introduction

Multiword expressions are groups of words which, taken together, can have unpredictable semantics. For example, the expression *part of speech* refers not to some aspect of speaking, but to the syntactic category of a word. If the expression is altered in some ways—*part of speeches*, *part of speaking*, *type of speech*—then the idiomatic meaning is lost. Other modifications, however, are permitted, as in the plural *parts of speech*. These characteristics make multiword expressions (MWEs) difficult to identify and classify. But if they can be identified, then the incorporation of MWE knowledge has been shown to improve task accuracy for a range of NLP applications

* Department of Computer Science. E-mail: `spenceg@stanford.edu`.
** Department of Linguistics. E-mail: `mcdm@stanford.edu`.
† Departments of Computer Science and Linguistics. E-mail: `manning@stanford.edu`.

Submission received: October 1, 2011; revised submission received: June 9, 2012; accepted for publication: August 3, 2012.

including dependency parsing (Nivre and Nilsson 2004), supertagging (Blunsom and Baldwin 2006), sentence generation (Hogan et al. 2007), machine translation (Carpuat and Diab 2010), and shallow parsing (Korkontzelos and Manandhar 2010).

The standard approach to MWE identification is ***n*-gram classification**. This technique is simple. Given a corpus, all *n*-grams are extracted, filtered using heuristics, and assigned feature vectors. Each coordinate in the feature vector is a real-valued quantity such as log likelihood or pointwise mutual information. A binary classifier is then trained to render a MWE/non-MWE decision. All entries into the 2008 MWE Shared Task (Evert 2008) utilized variants of this technique.

Broadly speaking, *n*-gram classification methods measure word co-occurrence. Suppose that a corpus contains more occurrences of *part of speech* than *parts of speech*. Surface statistics may erroneously predict that only the former is an MWE and the latter is not. More worrisome is that the statistics for the two *n*-grams are separate, thus missing an obvious generalization.

In this article, we show that statistical parsing models generalize more effectively over arbitrary-length multiword expressions. This approach has not been previously demonstrated. To show its effectiveness, we build two parsing models for MWE identification. The first model is based on a context-free grammar (CFG) with manual rule refinements (Klein and Manning 2003). This parser also includes a novel lexical model—the **factored lexicon**—that incorporates morphological features. The second model is based on **tree substitution grammar** (TSG), a formalism with greater strong generative capacity that can store larger structural tree fragments, some of which are lexicalized.

We apply the models to Modern Standard Arabic (henceforth MSA, or simply "Arabic") and French, two morphologically rich languages (MRLs). The lexical sparsity (in finite corpora) induced by rich morphology poses a particular challenge for *n*-gram classification. Relative to English, French has a richer array of morphological features—such as grammatical gender and verbal conjugation for aspect and voice. Arabic also has richer morphology including gender and dual number. It has pervasive verb-initial matrix clauses, although preposed subjects are also possible. For languages like these it is well known that constituency parsing models designed for English often do not generalize well. Therefore, we focus on the interplay among language, annotation choices, and parsing model design for each language (Levy and Manning 2003; Kübler 2005, inter alia), although our methods are ultimately very general.

Our modeling strategy for MWEs is simple: We mark them with flat bracketings in phrase structure trees. This representation implicitly assumes a locality constraint on idioms, an assumption with a precedent in linguistics (Marantz 1997, inter alia). Of course, it is easy to find non-local idioms that do not correspond to surface constituents or even contiguous strings (O'Grady 1998). Utterances such as *All hell seemed to break loose* and *The cat got Mary's tongue* are clearly idiomatic, yet the idiomatic elements are discontiguous. Our models cannot identify these MWEs, but then again, neither can *n*-gram classification. Nonetheless, many common MWE types like nominal compounds are contiguous and often correspond to constituent boundaries.

Consider again the phrasal compound *part of speech*,[1] which is non-compositional: The idiomatic meaning "syntactic category" does not derive from any of the component

1 It is common to hyphenate some nominal compounds, e.g., *part-of-speech*. This practice invites a *words-with-spaces* treatment of idioms. However, hyphens are inconsistently used in English. Hyphenation is more common in French, but totally absent in Arabic.

words. This non-compositionality affects the syntactic environment of the compound as shown by the addition of an attributive adjective:

(1) a. *Noun* is a part of speech.
 b. **Noun* is a big part of speech.
 c. **Noun* is a big part.

(2) a. Liquidity is a part of growth.
 b. Liquidity is a big part of growth.
 c. Liquidity is a big part.

In Example (1a) the copula predicate *part of speech* as a whole describes *Noun*. In Examples (1b) and (1c) *big* clearly modifies only *part* and the idiomatic meaning is lost. The attributive adjective cannot probe arbitrarily into the non-compositional compound. In contrast, Example (2) contains parallel data without idiomatic semantics. The conventional syntactic analysis of Example (2a) is identical to that of Example (1a) except for the lexical items, yet *part of growth* is not idiomatic. Consequently, many premodifiers are appropriate for *part*, which is semantically vacuous. In Example (2b), *big* clearly modifies *part*, and *of growth* is just an optional PP complement, as shown by Example (2c), which is still grammatical.

This article proposes different phrase structures for examples such as (1a) and (2a). Figure 1a shows a Penn Treebank (PTB) (Marcus, Marcinkiewicz, and Santorini 1993) parse of Example (1a), and Figure 1b shows the parse of a paraphrase. The phrasal compound *part of speech* functions syntactically like a single-word nominal like *category*, and indeed *Noun is a big category* is grammatical. Single-word paraphrasability is a common, though not mandatory, characteristic of MWEs (Baldwin and Kim 2010). Starting from the paraphrase parse, we create a representation like Figure (1c). The MWE is indicated by a label in the predicted structure, which is flat. This representation explicitly models the idiomatic semantics of the compound and is context-free, so we can build efficient parsers for it. Crucially, MWE identification becomes a by-product of parsing as we can trivially extract MWE spans from full parses.

We convert existing Arabic and French syntactic treebanks to the new MWE representation. With this representation, the TSG model yields the best MWE identification results for Arabic (81.9% F1) and competitive results for French (71.3%), even though its parsing results lag state-of-the-art probabilistic CFG (PCFG)-based parsers. The TSG model also learns human-interpretable MWE rules. The factored lexicon model with gold morphological annotations achieves the best MWE results for French (87.3% F1) and competitive results for Arabic (78.2% F1). For both languages the factored lexicon model also approaches state-of-the-art basic parsing accuracy.

The remainder of this article begins with linguistic background on common MWE types in Arabic and French (Section 2). We then describe two constituency parsing models that are tuned for MWE identification (Sections 3 and 4). These models are supervised and can be trained on existing linguistic resources (Section 5). We evaluate the models for both basic parsing and MWE identification (Section 6). Finally, we compare our results with a state-of-the-art *n*-gram classification system (Section 7) and to prior work (Section 8).

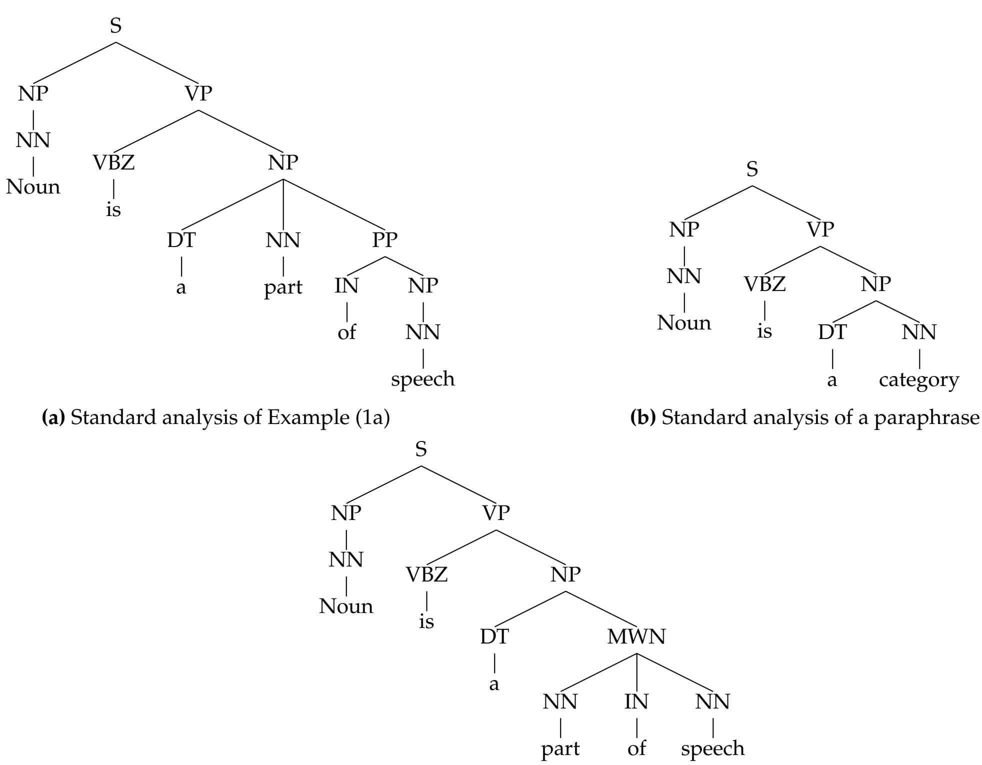
2. Multiword Expressions in Arabic and French

In this section we provide a general definition and taxonomy of MWEs. Then we discuss types of MWEs in Arabic and French.

2.1 Definition of Multiword Expressions

MWEs, a known nuisance for both linguistics and NLP, blur the lines between syntax and semantics. Jackendoff (1997, page 156) comments that MWEs "are hardly a marginal part of our use of language," and estimates that a native speaker knows at least as many MWEs as single words. A linguistically adequate representation for MWEs remains an active area of research, however. Baldwin and Kim (2010) define MWEs as follows:

Definition 1
Multiword expressions are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic, and/or statistical idiomaticity.



(a) Standard analysis of Example (1a)

(b) Standard analysis of a paraphrase

(c) MWE analysis used in this article

Figure 1
(a) A standard PTB parse of Example (1a). (b) The MWE *part of speech* functions syntactically like the ordinary nominal *category*, as shown by this paraphrase. (c) We incorporate the presence of the MWE into the syntactic analysis by flattening the tree dominating *part of speech* and introducing a new non-terminal label **multiword noun** (MWN) for the resulting span. The new representation classifies an MWE according to a global syntactic type and assigns a POS to each of the internal tokens. It makes no commitment to the internal syntactic structure of the MWE, however.

Table 1
Semi-fixed MWEs in French and English. The French adverb *à terme* ('in the end') can be modified by a small set of adjectives, and in turn some of these adjectives can be modified by an adverb such as *très* ('very'). Similar restrictions appear in English.

French				English			
à			terme	in the		near	term
à		court	terme	in the		short	term
à	très	court	terme	in the	very	short	term
à		moyen	terme	in the		medium	term
à		long	terme	in the		long	term
à	très	long	terme	in the	very	long	term

MWEs fall into four broad categories (Sag et al. 2002):

1. **Fixed**—do not allow morphosyntactic variation or internal modification (*in short*, *by and large*).
2. **Semi-fixed**—can be inflected or undergo internal modification (Table 1).
3. **Syntactically flexible**—undergo syntactic variation such as inflection (e.g., phrasal verbs such as *look up* and *write down*).
4. **Institutionalized phrases**—fully compositional phrases that are statistically idiosyncratic (*traffic light*, *Secretary of State*).

Statistical parsers are well-suited for coping with lexical, syntactic, and statistical idiomaticity across all four MWE classes. However, to our knowledge, we are the first to explicitly tune parsers for MWE identification.

2.2 Arabic MWEs

The most recent and most relevant work on Arabic MWEs was by Ashraf (2012), who analyzed an 83-million-word Arabic corpus. He developed an empirical taxonomy of six MWE types, which correspond to syntactic classes. The syntactic class is defined by the projection of the purported syntactic head of the MWE. MWEs are further subcategorized by observed POS sequences. For some of these classes, the syntactic distinctions are debatable. For example, in the verb-object idiom ضرب عصفورين بحجر *Daraba ^{c}Sfuurayn bi-Hajar* ('he killed two birds with one stone')[2] the composition of عصفورين ('two birds') with حجر ('stone') is at least as important as composition with the verb ضرب ('he killed'), yet Ashraf (2012) classifies the phrase as a verbal idiom.

The corpus in our experiments only marks three of the six Arabic MWE classes:

Nominal idioms (MWN) consist of proper nouns (Example 3a), noun compounds (Example 3b), and construct NPs (Example 3c). MWNs typically correspond to NP bracketings:

(3) a. N N: ابو ظبي *abuu Dabii* ('Abu Dhabi')

2 For each Arabic example in this work, we provide native script, transliterations in italics according to the phonetic scheme in Ryding (2005), and English translations in single quotes.

b. D+N D+N: العناية الفائقة *al-^{c}naaya al-faa$^\circ$qa* ('intensive care unit')

c. N D+N: كرة القدم *kura al-qudum* ('soccer')

Prepositional idioms (MWP) include PPs that are commonly used as discourse connectives (Example 4a), function like adverbials in English (Example 4b), or have been institutionalized (Example 4c). These MWEs are distinguished by a prepositional syntactic head:

(4) a. P D+N: حتى الآن *Hataa al-aan* ('until now')

b. P+N: بعنف *bi-^{c}nf* ('violently')

c. P+D+N D+N: بالتوقيت المحلي *bi-al-twqiit al-maHalii* ('local time')

Adjectival idioms (MWA) are typically the so-called "false" *iDaafa* constructs in which the first term is an adjective that acts as a modifier of some other noun. These constructs often correspond to a hyphenated modifier in English such as Examples (5a) and (5b). Less frequent are coordinated adjectives that have been institutionalized such as Examples (5c) and (5d):

(5) a. A D+N: رفيعة المستوى *rafii^{c}a al-mustuuaa* ('high-level')

b. A D+N: سوفياتية الصنع *swfiiaatiia al-Sana^c* ('Soviet-made')

c. D+A C+D+A: الشقيقة والصديقة *al-shaqiiqa w-al-Sadiiqa* ('neighborly')

d. D+A C+D+A: البرية والبحرية *al-bariia w-al-baHariia* ('land and sea')

These idiom types usually do not cross constituent boundaries, so constituency parsers are well suited for modeling them. The other three classes of Ashraf (2012)—verb-subject, verbal, and adverbial—tend to cross constituent boundaries, so they are difficult to represent in a PTB-style treebank. Dependency representations may be more appropriate for these idiom classes.

2.3 French MWEs

In French, there is a lexicographic tradition of compiling MWE lists. For example, Gross (1986) shows that whereas French dictionaries contain about 1,500 single-word adverbs there are over 5,000 multiword adverbs. MWEs occur in every part of speech (POS) category (e.g., noun *trousse de secours* ('first-aid kit'); verb *faire main-basse* [do hand-low] ('seize'); adverb *comme dans du beurre* [as in butter] ('easily'); adjective *à part entière* ('wholly')).

Motivated by the prevalence of MWEs in French, Gross (1984) developed a linguistic theory known as Lexicon-Grammar. In this theory, MWEs are classified according to their global POS tags (noun, verb, adverb, adjective), and described in terms of the sequence of the POS tags of the words that constitute the MWE (e.g., "N de N" *garde d'enfant* [guard of child] ('daycare'), *pied de guerre* [foot of war] ('at the ready')) (Gross 1986). In other words, MWEs are represented by a flat structure. The Lexicon-Grammar distinguishes between units that are fixed and have to appear as is (*en tout et pour tout* [in all and for all] ('in total')) and units that accept some syntactic variation such as admitting the insertion of an adverb or adjective, or the variation of one of the words in the expression (e.g., a possessive as in *from the top of one's hat*). It also notes whether the MWE displays some selectional preferences (e.g., it has to be preceded by a verb or by an adjective).

We discuss three of the French MWE categories here, and list the rest in Appendix A.

Nominal idioms (MWN) consist of proper nouns (Example (6a)), foreign common nouns (6b), and common nouns. The common nouns appear in several syntactically regular sequences of POS tags (Example (7)). Multiword nouns allow inflection (singular vs. plural) but no insertion:

(6) a. *London Sunday Times, Los Angeles*

b. *week - end, mea culpa, joint - venture*

(7) a. N A: *corps médical* ('medical staff'), *dette publique* ('public debt')

b. N P N: *mode d'emploi* ('instruction manual')

c. N N: *numéro deux* ('number two'), *maison mère* [house mother] ('headquarters'), *grève surprise* ('sudden strike')

d. N P D N: *impôt sur le revenu* ('income tax'), *ministre de l'économie* ('finance minister')

Adjectival idioms (MWA) appear with different POS sequences (Example (8)). They include numbers like *vingt et unième* ('21st'). Some MWAs allow internal variation. For example, some adverbs or adjectives can be added to both examples in (8b) (*à <u>très</u> haut risque*, *de <u>toute</u> dernière minute*):

(8) a. P N: *d'antan* [from before] ('old'), *en question* ('under discussion')

b. P A N: *à haut risque* ('high-risk'), *de dernière minute* [from the last minute] ('at the eleventh hour')

c. A C A: *pur et simple* [pure and simple] ('straightforward'), *noir et blanc* ('black and white')

Verbal idioms (MWV) allow number and tense inflections (Example (9)). Some MWVs containing a noun or an adjective allow the insertion of a modifier (e.g., *donner grande satisfication* ('give great satisfaction')), whereas others do not. When an adverb intervenes between the main verb and its complement, the two parts of the MWV may be marked discontinuously (e.g., [$_{MWV}$ [$_{V}$ prennent]] [^{ADV} déjà] [$_{MWV}$ [$_{P}$ en] [$_{N}$ cause]] ('already take into account')):

(9) a. V N: *avoir lieu* ('take place'), *donner satisfaction* ('give satisfaction')

b. V P N: *mettre en place* ('put in place'), *entrer en vigueur* ('to come into effect')

c. V P ADV: *mettre à mal* [put at bad] ('harm'), *être à même* [be at same] ('be able')

d. V D N P N: *tirer la sonnette d'alarme* ('ring the alarm bell'), *avoir le vent en poupe* ('to have the wind astern')

Both Gross (1986) and Ashraf (2012) classify MWEs according to global syntactic role and internal POS sequence. In a constituency tree, these two features can be modeled

Table 2
French grammar development. Incremental effects on grammar size and labeled F1 for each of the manual grammar features (development set, sentences $\leq$ 40 words). The baseline is a parent-annotated grammar. The features tradeoff between maximizing two objectives: overall parsing F1 and MWE F1.

Feature	States	Tags	Parse F1	ΔF1	MWE F1
–	4,128	32	77.3		60.7
tagPA	4,360	264	78.4	+1.1	71.4
splitPUNC	4,762	268	78.8	+0.4	71.1
markDe	4,882	284	79.8	+1.0	71.6
markP	4,884	286	79.9	+0.1	71.5
MWADVtype1	4,919	286	79.9	+0.0	71.8
MWADVtype2	4,970	286	79.9	+0.0	71.7
MWNtype1	5,042	286	80.0	+0.1	71.9
MWNtype2	5,098	286	79.9	−0.1	71.9

by a span over the MWE composed of a phrasal label indicating the MWE type and pre-terminal labels indicating the internal POS sequence. MWE identification then becomes a trivial process of extracting such subtrees from full parses.

3. Context-Free Parsing Model: Stanford Parser

In this section and Section 4, we describe constituency parsing models that will be tuned for MWE identification. The algorithmic details of the parsing models may seem removed from multiword expressions, but this is by design. MWEs are encoded in the syntactic representation, allowing the model designer to focus on learning that representation rather than trying to model semantic phenomena directly.

The Stanford parser (Klein and Manning 2003) is a product model that combines the outputs of a manually refined PCFG with an arc-factored dependency parser. Adapting the Stanford parser to a new language requires: (1) feature engineering for the PCFG grammar, (2) specification of head-finding rules for extracting dependencies, and (3) development of an unknown word model.[3]

After adapting the basic parser, we develop a novel lexical model, which we call a **factored lexicon**. The factored lexicon incorporates morphological information that is predicted by a separate morphological analyzer.

3.1 Grammar Development

Grammar features consist of manual splits of labels in the training data (e.g., marking base NPs with the rich label "NP-base"). These features were tuned on a development set. Some of them have linguistic interpretations, whereas others (e.g., punctuation splitting) have only empirical justification.

French Grammar Features. Table 2 lists the category splits used in our grammar. Most of the features are POS splits as many phrasal tag splits did not improve accuracy. This result may be due to the flat annotation scheme of the FTB.

3 The Stanford parser code, head-finding rules, and trained models are available at `http://nlp.stanford.edu/software/lex-parser.shtml`.

Parent annotation of POS tags captures information about the external context. For example, prepositions (P) can introduce a prepositional phrase (PP) or an infinitival complement (VPinf), but some prepositions will uniquely appear in one context and not the other (e.g., *sur* ('on') will only occur in a PP environment). The **tagPA** provides this kind of distribution. We also split punctuation tags (**splitPUNC**) into equivalence classes similar to those present in the PTB.

We tried different features to mark the context of prepositions. **markP** identifies prepositions which introduce PPs modifying a noun (NP). Marking other kinds of prepositional modifiers (e.g., verb) did not help. The feature **markDe** the preposition *de* and its variants (*du*, *des*, *d'*), which are very frequent and appear in many contexts.

The features that help MWE F1 depend on idiom frequency. We mark MWADVs under S nodes (**MWADVtype1**), and those with POS sequences that occur more than 500 times ("P N" – *en jeu*, *à peine*, or "P D N" *dans l'immédiat*, *à l'inverse*) (**MWADVtype2**). Similarly, we mark MWNs that occur more than 600 times (e.g., "N P N" and "N N") (**MWNtype1** and **MWNtype2**).

Arabic Grammar Features. The Arabic grammar features come from Green and Manning (2010), which contains an ablation study similar to Table 2. We added one additional feature, **markMWEPOS**, which marks POS tags dominated by MWE phrasal categories.

3.2 Head-Finding Rules

For Arabic, we use the head-finding rules from Green and Manning (2010). For French, we use the head-finding rules of Dybro-Johansen (2004), which yielded an approximately 1% development set improvement over those of Arun (2004).

3.3 Unknown Word Models

For both languages, we create simple unknown word models that substitute word signatures for rare and unseen word types. The signatures are generated according to the features in Table 3. For tag t and signature s, the signature parameters $p(t|s)$ are estimated after collecting counts for 50% of the training data. Then $p(s|t)$ is computed via Bayes rule with a flat Dirichlet prior.

Table 3
Unknown word model features for Arabic and French.

Arabic Lexical Features	French Lexical Features
▷ Presence of the determiner ال *al*	▷ Nominal, adjectival, verbal, adverbial, and plural suffixes
▷ Contains digits or punctuation	▷ Contains digits or punctuation
▷ Ends with the feminine affix ة *ah*	▷ Is capitalized (except the first word in a sentence), or consists entirely of capital letters
▷ Various verbal (e.g., ت *t*, وا *waa*, ون *uun*) and adjectival suffixes (e.g., ية *iiah*, ي *ii*)	▷ If none of the above, deterministically extract one- and two-character suffixes

3.4 Factored Lexicon with Morphological Features

We will apply our models to Arabic and French, yet we have not dealt with the lexical sparsity induced by rich morphology (see Table 5 for a comparison to English). One way to combat sparsity is to parse a **factored** representation of the terminals, where factors might be the word form, the lemma, or grammatical features such as gender, number, and person (ϕ features) (Bilmes and Kirchoff 2003; Koehn and Hoang 2007, inter alia).

The basic parser lexicon estimates the generative probability of a word given a tag $p(w|t)$ from word/tag pairs observed in the training set. Additionally, the lexicon includes parameter estimates $p(t|s)$ for unknown word signatures s produced by the unknown word models (see Section 3.3). At parsing time, the lexicon scores each input word type w according to its observed count in the training set $c(w)$. We define the unsmoothed and smoothed parameter estimates:

$$p(t|w) = \frac{c(t,w)}{c(w)} \tag{1}$$

$$p_{smooth}(t|w) = \frac{c(t,w) + \alpha p(t|s)}{c(w) + \alpha} \tag{2}$$

We then compute the desired parameter $p(w|t)$ as

$$p(w|t) = \begin{cases} \frac{p(t|w)p(w)}{p(t)} & \text{if } c(w) > \beta \\ \frac{p_{smooth}(t|w)p(w)}{p(t)} & \text{if } c(w) > 0 \\ \frac{p(t|s)p(s)}{p(t)} & \text{otherwise} \end{cases} \tag{3}$$

We found that $\alpha = 1.0$ and $\beta = 100$ worked well on both development sets.

In the factored lexicon, each token has an associated morphological analysis m, which is a string describing various grammatical features (e.g., tense, voice, definiteness). Instead of generating terminals alone, we generate the word and morphological analysis using a simple product:

$$p(w,m|t) = p(w|t)p(m|t) \tag{4}$$

where $p(m|t)$ is estimated using exactly the same procedure as the lexical insertion probability $p(w|t)$. Because there are only a few hundred unique $\langle t, m \rangle$ tuples in the training data for each language, we tend to get sharper parameter estimates, namely, we usually estimate $p(t|m)$ directly as in Equation (1). Moreover, at test time, even if the word type w is unknown, the associated morphological analysis m is almost always known, providing additional evidence for tagging.

We also experimented with an additional lemma factor, but found that it did not improve accuracy. We thus excluded the lemma factor from our experiments.

For words that have been observed with only one tagging, the factored lexicon is clearly redundant. Consider, however, the case of the Arabic triliteral قتل *qtl* which, in unvocalized text, can be either a verb meaning "he killed" or a nominal meaning "murder, killing." If قتل appears as a verb, and we include the tense feature in the morphological analysis, then all associated nominal tags (e.g., NN) will be assigned zero probability because nominals never carry tense in the training data.

4. Fragment Parsing Model: Dirichlet Process Tree Substitution Grammars

For our task, a shortcoming of CFG-based grammars is that they do not explicitly capture idiomatic usage. For example, consider the two utterances:

(10) a. He kicked the bucket.

b. He kicked the pail.

Unless horizontal markovization is applied, PCFGs generate words independently. Consequently, no phrasal rule parameter in the model differentiates between Examples (10a) and (10b). Recall, however, that in our representation, Example (10a) should receive a flat analysis as MWV, whereas Example (10b) should have a conventional analysis of the transitive verb *kicked* and its two arguments.

TSGs are weakly equivalent to CFGs, but with greater strong generative capacity (Joshi and Schabes 1997). TSGs can store lexicalized **tree fragments** as rules. Consequently, if we have seen [$_{\text{MWV}}$ *kicked the bucket*] several times before, we can store that whole lexicalized fragment in the grammar.

We consider the non-parametric probabilistic TSG (PTSG) model of Cohn, Goldwater, and Blunsom (2009) in which tree fragments are drawn from a Dirichlet process (DP) prior.[4] The DP-TSG can be viewed as a data-oriented parsing (DOP) model (Scha 1990; Bod 1992) with Bayesian parameter estimation. A PTSG is a 5-tuple $\langle V, \Sigma, R, \diamondsuit, \theta \rangle$ where $c \in V$ are non-terminals; $t \in \Sigma$ are terminals; $e \in R$ are **elementary trees**;[5] $\diamondsuit \in V$ is a unique start symbol; and $\theta_{c,e} \in \theta$ are parameters for each tree fragment. A PTSG derivation is created by successively applying the substitution operator to the leftmost **frontier node** (denoted by c^+). All other nodes are **internal** (denoted by c^-).

In the supervised setting, DP-TSG grammar extraction reduces to a segmentation problem. We have a treebank T that we segment into the set R, a process that is modeled with Bayes' rule:

$$p(R \mid T) \propto p(T \mid R)\, p(R) \tag{5}$$

Because the tree fragments completely specify each tree, $p(T \mid R)$ is either 0 or 1, so all work is performed by the prior over the set of elementary trees.

The DP-TSG contains a DP prior for each $c \in V$ and generates a tree fragment e rooted at non-terminal c according to:

$$\theta_c | c, \alpha_c, P_0(\cdot|c) \sim DP(\alpha_c, P_0)$$
$$e | \theta_c \sim \theta_c$$

4 Similar models were developed independently by O'Donnell, Tenenbaum, and Goodman (2009) and Post and Gildea (2009).

5 We use the terms *tree fragment* and *elementary tree* interchangeably.

Table 4
DP-TSG notation. For consistency, we largely follow the notation of Liang, Jordan, and Klein (2010).

α_c	DP concentration parameter for each non-terminal type $c \in V$
$P_0(e\|c)$	CFG base distribution
$\mathbf{x}$	Set of all non-terminal nodes in the treebank
$\mathcal{S}$	Set of sampling sites (one for each $x \in \mathbf{x}$)
S	A block of sampling sites, where $S \subseteq \mathcal{S}$
$\mathbf{b} = \{b_s\}_{s \in \mathcal{S}}$	Binary variables to be sampled ($b_s = 1$ for frontier nodes)
$\mathbf{z}$	Latent state of the segmented treebank
m	Number of sites $s \in S$ s.t. $b_s = 1$
$\mathbf{n} = \{n_{c,e}\}$	Sufficient statistics of $\mathbf{z}$
$\Delta n^{S:m}$	Change in counts by setting m sites in S

Table 4 defines notation. The data likelihood is given by the latent state $\mathbf{z}$ and the parameters θ: $p(\mathbf{z}|\theta) = \prod_{z \in \mathbf{z}} \theta_{c,e}^{n_{c,e}(\mathbf{z})}$. Integrating out the parameters, we have:

$$p(\mathbf{z}) = \prod_{c \in V} \frac{\prod_e (\alpha_c P_0(e|c))^{\overline{n_{c,e}(\mathbf{z})}}}{\alpha_c^{\overline{n_{c,\cdot}(\mathbf{z})}}} \tag{6}$$

where $x^{\overline{n}} = x(x+1)\ldots(x+n-1)$ is the rising factorial.

Base Distribution. The base distribution P_0 is the same maximum likelihood PCFG used in the Stanford parser.[6] After applying the manual grammar features, we perform simple right binarization, collapse unary rules, and replace rare words with their signatures (Petrov et al. 2006).

For each non-terminal type c, we learn a stop probability $q_c \sim \text{Beta}(1, 1)$. Under P_0, the probability of generating a tree fragment $A^+ \rightarrow B^- \, C^+$ composed of non-terminals is

$$P_0(A^+ \rightarrow B^- \, C^+) = p_{\text{MLE}}(A \rightarrow B\, C) q_B (1 - q_C) \tag{7}$$

Unlike Cohn, Goldwater, and Blunsom (2009), we penalize lexical insertion:

$$P_0(c \rightarrow t) = p_{\text{MLE}}(c \rightarrow t) p(t) \tag{8}$$

where $p(t)$ is equal to the MLE unigram probability of t in the treebank. Lexicalizing a rule makes it very specific, so we generally want to avoid lexicalization with rare words. Empirically, we found that this penalty reduces overfitting.

Type-Based Inference Algorithm. To learn the parameters θ we use the collapsed, block Gibbs sampler of Liang, Jordan, and Klein (2010). We sample binary variables b_s associated with each sampling site s in the treebank. The key idea is to select a block

6 The Stanford parser is a product model which scores parses with both a dependency grammar and a PCFG. We extract the TSG from the manually split PCFG only. Bansal and Klein (2010) also experimented with manual grammar features in an all-fragments (parametric) TSG for English.

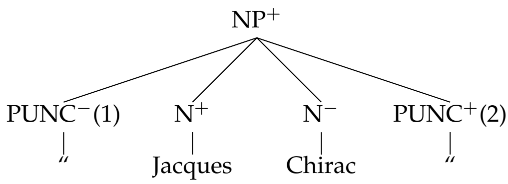


Figure 2
Example of two conflicting sites of the same type in a training tree. Define the type of a site $t(z, s) \stackrel{\text{def}}{=} (\Delta n^{s:0}, \Delta n^{s:1})$. Sites (1) and (2) have the same type because $t(z, s_1) = t(z, s_2)$. The two sites conflict, however, because the probabilities of setting b_{s_1} and b_{s_2} both depend on counts for the tree fragment rooted at NP. Consequently, sites (1) and (2) are not exchangeable: The probabilities of their assignments depend on the order in which they are sampled.

of exchangeable sites S of the same **type** that do not **conflict** (Figure 2). Because the sites in S are exchangeable, we can set b_S randomly if we know m, the number of sites with $b_s = 1$. This algorithm is not a contribution of this article, so we refer the interested reader to Liang, Jordan, and Klein (2010) for further details.

After each Gibbs iteration, we sample each stop probability q_c directly using binomial-Beta conjugacy. We also infer the DP concentration parameters α_c with the auxiliary variable procedure of West (1995).

Decoding. To decode, we first create a maximum a posterior (MAP) grammar in which tree fragments have fixed estimates according to a single sample from the DP-TSG:

$$\theta_{c,e} = \frac{n_{c,e}(z) + \alpha_c P_0(e|c)}{n_{c,}(z) + \alpha_c} \tag{9}$$

This MAP grammar has an infinite rule set, however, because elementary trees with zero count in n have some residual probability under P_0. We discard all zero-count trees except for the zero-count CFG rules in P_0. Scores for these rules follow from Equation (9) with $n_{c,e}(z) = 0$. This grammar represents most of the probability mass and permits inference using dynamic programming (Cohn, Goldwater, and Blunsom 2009).

Because the derivations of a TSG are context-free (Vijay-Shanker and Weir 1993), we can form a CFG of the derivation sequences and use a synchronous CFG to translate the most probable CFG parse to its TSG derivation. Consider a unique tree fragment e_i rooted at c_j with frontier γ, which is a sequence of terminals and non-terminals. We encode this fragment as an SCFG rule of the form

$$[c_j \rightarrow \gamma \, , \, c_j \rightarrow i, c_k, c_l, \ldots] \tag{10}$$

where $c_k, c_l, \ldots$ is a finite-length sequence of the non-terminal frontier nodes in γ.[7] The SCFG translates the input string to a sequence of tree fragment indices. Because the TSG substitution operator applies to the leftmost frontier node, the best TSG parse can be deterministically recovered from the sequence of indices.

7 This formulation is due to Chris Dyer.

Table 5
Gross corpus statistics for the pre-processed corpora used to train and evaluate our models. We compare to the WSJ section of the PTB: train (Sections 02–21); dev. (Section 22); test (Section 23). Due to its flat annotation style, the FTB sentences have fewer constituents per sentence. In the ATB, morphological variation accounts for the high proportion of word types to sentences.

		ATB	FTB	WSJ
Train	#sentences	18,818	13,448	39,832
	#tokens	597,933	397,917	950,028
	#word types	37,188	26,536	44,389
	#POS types	32	30	45
	#phrasal types	31	24	27
	avg. length	31.8	29.6	23.9
Dev.	#sentences	2,318	1,235	1,700
	#tokens	70,656	38,298	40,117
	#word types	12,358	6,794	6,840
	avg. length	30.5	31.0	23.6
	OOV rate	15.6%	17.8%	12.8%
Test	#sentences	2,313	1,235	2,416
	#tokens	70,065	37,961	56,684

The SCFG formulation has a practical benefit: We can take advantage of the heavily optimized SCFG decoders for machine translation. We use `cdec` (Dyer et al. 2010) to find the Viterbi derivation for each input string.

5. Training Data and Morphological Analyzers

We have described two supervised parsing models for Arabic and French. Now we show how to construct MWE-aware training resources for them.

The corpora used in our experiments are the Penn Arabic Treebank (ATB) (Maamouri et al. 2004) and the French Treebank (FTB) (Abeillé, Clément, and Kinyon 2003). Prior to parsing, both treebanks require significant pre-processing, which we perform automatically.[8] Because parsing evaluation metrics are sensitive to the terminal/non-terminal ratio (Rehbein and van Genabith 2007), we only remove non-terminal labels in the case of unary rewrites of the same category (e.g., NP $\rightarrow$ NP) (Johnson 1998). Table 5 compares the pre-processed corpora with the WSJ section of the PTB. Appendix C compares the annotation consistency of the ATB, FTB, and WSJ.

5.1 Arabic Treebank

We work with parts 1–3 (newswire) of the ATB,[9] which contain documents from three different news agencies. In addition to phrase structure markers, each syntactic tree also contains per-token morphological analyses.

8 Tree manipulation is automated with Tregex/Tsurgeon (Levy and Andrew 2006). Our pre-processing package is available at `http://nlp.stanford.edu/software/lex-parser.shtml`.
9 LDC catalog numbers: LDC2008E61, LDC2008E62, and LDC2008E22.

Table 6
Frequency distribution of the MWE types in the ATB and FTB training sets.

Categories		ATB		FTB	
MWN	*noun*	6,975	91.6%	9,680	49.7%
MWP	*prep.*	623	8.18%	3,526	18.1%
MWA	*adj.*	18	0.24%	324	1.66%
MWPRO	*pron.*	–	–	266	1.37%
MWC	*conj.*	–	–	814	4.18%
MWADV	*adverb*	–	–	3,852	19.8%
MWV	*verb*	–	–	585	3.01%
MWD	*det.*	–	–	328	1.69%
MWCL	*clitic*	–	–	59	0.30%
MWET	*foreign*	–	–	24	0.12%
MWI	*interj.*	–	–	4	0.02%
Total		7,616		19,462	

Tokenization/Segmentation. We retained the default ATB clitic segmentation scheme.

Morphological Analysis. The ATB contains gold per-token morphological analyses, but no lemmas.

Tag Sets. We used the POS tag set described by Kulick, Gabbard, and Marcus (2006). We previously showed that the "Kulick" tag set is very effective for basic Arabic parsing (Green and Manning 2010).

MWE Tagging. The ATB does not mark MWEs. Therefore, we merged an existing Arabic MWE list (Attia et al. 2010b) with the constituency trees.[10] For each string from the MWE list that was bracketed in the treebank, we flattened the structure over the MWE span and added a non-terminal label according to the MWE type (Table 6). We ignored MWE strings that crossed constituent boundaries.

Orthographic Normalization. Orthographic normalization has a significant impact on parsing accuracy. We remove all diacritics, instances of *taTwiil*,[11] and pro-drop markers. We also applied *alif* normalization[12] and mapped punctuation and numbers to their Latin equivalents.

Corpus Split. We divided the ATB into training/development/test sections according to the split prepared by Mona Diab for the 2005 Johns Hopkins workshop on parsing Arabic dialects (Rambow et al. 2005).[13]

10 The list of 30,277 distinct MWEs is available at: `http://sourceforge.net/projects/arabicmwes/`.

11 *taTwiil* (ـ) is an elongation character for justifying text. It has no morphosyntactic function or phonetic realization.

12 Variants of *alif* [ا,أ,إ,آ] are inconsistent in Arabic text.

13 The corpus split is available at: `http://nlp.stanford.edu/projects/arabic.shtml`.

5.2 French Treebank

The FTB[14] contains phrase structure trees with morphological analyses and lemmas. In addition, the FTB explicitly annotates MWEs. POS tags for MWEs are given not only at the MWE level, but also internally: Most tokens that constitute an MWE also have a POS tag. Our FTB pre-processing is largely consistent with Lexicon-Grammar, which defines MWE categories based on the global POS.

Tokenization/Segmentation. We changed the default tokenization for numbers by fusing adjacent digit tokens. For example, *500 000* is tagged as an MWE composed of two words *500* and *000*. We made this *500000* and removed the MWE POS. We also merged numbers like "17,9".

Morphological Analysis. The FTB provides both gold morphological analyses and lemmas for 86.6% of the tokens. The remaining tokens lack morphological analyses, and in many cases basic parts of speech. We restored the basic parts of speech by assigning each token its most frequent POS tag elsewhere in the treebank.[15] This technique was too coarse for missing morphological analyses, which we left empty.

Tag Sets. We transformed the raw POS tags to the CC tag set (Crabbé and Candito 2008), which is now the standard tag set in the French parsing literature. The CC tag set includes WH markers and verbal mood information.

MWE Tagging. We added the 11 MWE labels shown in Table 6. We mark MWEs with a flat bracketing in which the phrasal label is the MWE-level POS tag with an MW prefix, and the preterminals are the internal POS tags for each terminal. The resulting POS sequences are not always unique to MWEs: They appear in abundance elsewhere in the corpus. Some MWEs contain normally ungrammatical POS sequences, however (e.g., adverb *à la va vite* ('in a hurry'): P D V ADV [at the goes quick]), and some words appear only as part of an MWE, such as *insu* in *à l'insu de* ('to the ignorance of'). We also found that 36 MWE spans still lacked a global POS. To restore these labels, we assigned the most frequent label for that internal POS sequence elsewhere in the corpus.

Corpus Split. We used the 80/10/10 split described by Crabbé and Candito (2008). They used a previous release of the treebank with 12,531 trees. Subsequently, 3,391 trees were added to the FTB. We appended these extra trees to the training set, thus preserving the original development and test sets.

5.3 Morphological Analysis for Arabic and French

The factored lexicon requires predicted per-token morphological analyses at test time. We used separately trained, language-specific tools to obtain these analyses (Table 7).

14 Version from June 2010. We used the subset of the FTB with functional annotations, not for those annotations but because this subset is known to be more consistently annotated. Appendix B compares our pre-processed version of the FTB to other versions in prior work.

15 Seventy-three of the unlabeled word types did not appear elsewhere in the treebank. All but 11 of these were nouns. We manually assigned the correct tags, but we would not expect a negative effect by deterministically labeling all of them as nouns.

Table 7
Linguistic resources required by the factored lexicon. Equivalent resources for Arabic and French do not presently exist. The ATB lacks gold lemmas and a French morphological ranker equivalent to MADA—which can produce the full set of morphosyntactic features specified in the ATB—has not been developed. Morfette is effectively a discriminative classifier that treats analyses as atomic labels, whereas MADA utilizes a morphological generator.

	Arabic (ATB)	**French (FTB)**
Gold Morphological Features	Gender, Number, Tense, Person, Mood, Voice, Definiteness	Gender, Number, Tense, Person
Gold Lemmas	×	✓
Morphological Analyzer	✓ (SAMA)	×
Morphological Ranker	✓ (MADA)	✓ (Morfette)
Lemmatizer	✓ (MADA)	✓ (Morfette)

Arabic. The morphological analyses in the ATB are human-selected outputs of the Standard Arabic Morphological Analyzer (SAMA),[16] a deterministic system that relies on manually compiled linguistic dictionaries. The latest version of SAMA has complete lexical coverage of the ATB, thus it does not encounter unseen word types at test time. To rank the output of SAMA, we use MADA (Habash and Rambow 2005),[17] which makes predictions based on an ensemble of support vector machine (SVM) classifiers.

French. The FTB includes morphological analyses for gender, number, person, tense, type of pronouns (relative, reflexive, interrogative), type of adverbs (relative or interrogative), and type of nouns (proper vs. common noun). Morfette (Chrupala, Dinu, and van Genabith 2008) has been used in previous FTB parsing experiments (Candito and Seddah 2010; Seddah et al. 2010) to predict these features in addition to lemmas. Morfette is a discriminative sequence classifier that relies on lexical and greedy left context features. Because Morfette lacks a morphological generator like SAMA, however, it is effectively a tagger that must predict a very large tag set. We trained Morfette on our split of the FTB and evaluated accuracy on the development set: 88.3% (full morphological tagging); 95.0% (lemmatization); and 86.5% (full tagging and lemmatization).[18]

6. Experiments

For each language, we ran two experiments: standard parsing and MWE identification. The evaluation included the Stanford, Stanford+factored lexicon, and DP-TSG models.

All experiments used gold tokenization/segmentation. Unlike the ATB, the FTB does not contain the raw source documents, so we could not start from raw text for both

16 LDC catalog number LDC2010L01.

17 We used version 3.1. According to the user manual, the training set for the distributed models overlaps with our ATB development and test sets. Training scripts/procedures are not distributed with MADA, however.

18 Morfette training settings: 10 tag and 3 lemma training iterations. We excluded punctuation tokens from the morphological tagging evaluation because our parsers split punctuation deterministically.

languages. We previously showed that segmentation errors decrease Arabic parsing accuracy by about 2.0% F1 (Green and Manning 2010).

Morphological analysis accuracy was another experimental resource asymmetry between the two languages. The morphological analyses were obtained with significantly different tools: in Arabic, we had a morphological generator/ranker (MADA), whereas for French we had only a discriminative classifier (Morfette). Consequently, French analysis quality was lower (Section 5.3).

6.1 Standard Parsing Experiments

Baselines. We included two parsing baselines: a parent-annotated PCFG (PAPCFG) and a PCFG with the grammar features in the Stanford parser (SplitPCFG). The PAPCFG is the standard baseline for TSG models (Cohn, Goldwater, and Blunsom 2009).

Berkeley Parser. We previously showed optimal Berkeley parser (Petrov et al. 2006) parameterizations for both the Arabic (Green and Manning 2010) and French (Green et al. 2011) data sets.[19] For Arabic, our pre-processing and parameter settings significantly increased the best-published Berkeley ATB baseline. Others had used the Berkeley parser for French, but on an older revision of the FTB. To our knowledge, we are the first to use the Berkeley parser for MWE identification.

Factored Lexicon Features. We selected features for the factored lexicon on the development sets. For Arabic, we used gender, number, tense, mood, and definiteness. For French, we used the grammatical and syntactic features in the CC tag set in addition to grammatical number. For the experiments in which we evaluated with predicted morphological analyses, we also trained the parser on predicted analyses.

Evaluation Metrics. We report three evaluation metrics. **Evalb** is the standard labeled precision/recall metric.[20] **Leaf Ancestor** measures the cost of transforming guess trees to the reference (Sampson and Babarczy 2003), and is less biased against flat treebanks like the FTB (Rehbein and van Genabith 2007). The Leaf Ancestor score ranges from 0 to 1 (higher is better). We report micro-averaged (Corpus) and macro-averaged (Sent.) scores. Finally, **EX%** is the percentage of perfectly parsed sentences according to Evalb.

Sentence Lengths. We report results for sentences of lengths $\leq$ 40 words. This cutoff accounts for similar proportions of the ATB and FTB. The DP-TSG grammar extractor produces very large grammars for Arabic,[21] and we found that the grammar constant was too large for parsing all sentences. For example, the ATB development set contains a sentence that is 268 tokens long.

19 Berkeley training settings: right binarization, no parent annotation, and six split-merge cycles. Results are the average of three runs in which the random number generator was seeded with the system time.

20 Available at `http://nlp.cs.nyu.edu/evalb/` (v.20080701). We used a Java re-implementation included in the Stanford parser distribution that is compatible with the reference implementation.

21 Average DP-TSG grammar sizes: Arabic, 89,003 rules; French, 46,515 rules.

Table 8
Arabic standard parsing experiments (test set, sentences ≤ 40 words). SplitPCFG is the same grammar used in the Stanford parser, but without the dependency model. FactLex uses basic POS tags predicted by the parser and morphological analyses from MADA. FactLex* uses gold morphological analyses. Berkeley and DP-TSG results are the average of three independent runs.

Arabic	**Leaf Ancestor**		**Evalb**			
	Sent.	Corpus	LP	LR	F1	EX%
PAPCFG	0.777	0.745	69.5	64.6	66.9	12.9
SplitPCFG	0.821	0.797	75.6	73.4	74.5	17.8
Berkeley	**0.865**	**0.853**	**83.3**	**82.7**	**83.0**	**24.0**
DP-TSG	0.822	0.800	75.5	75.4	75.4	17.7
Stanford	0.851	0.835	81.3	80.7	81.0	23.5
Stanford+FactLex	0.849	0.835	81.2	80.8	81.0	22.8
Stanford+FactLex*	0.852	0.837	81.8	81.3	81.5	**24.0**

Results. Tables 8 and 9 show Arabic and French parsing results, respectively. For both languages, the Berkeley parser produces the best results in terms of Evalb F1. The gold factored lexicon setting compares favorably in terms of exact match.

6.2 MWE Identification Experiments

The predominant approach to MWE identification is the combination of lexical association measures (surface statistics) with a binary classifier (Pecina 2010). A state-of-the-art, language-independent package that implements this approach for higher order n-grams is `mwetoolkit` (Ramisch, Villavicencio, and Boitet 2010).

`mwetoolkit` *Baseline.* We configured `mwetoolkit` with the four standard lexical features: the maximum likelihood estimator, Dice's coefficient, pointwise mutual information, and Student's t-score. We also included POS tags predicted by the Stanford tagger (Toutanova et al. 2003). We filtered the training instances by removing unigrams and

Table 9
French standard parsing experiments (test set, sentences ≤ 40 words). FactLex uses basic POS tags predicted by the parser and morphological analyses from Morfette. FactLex* uses gold morphological analyses.

French	**Leaf Ancestor**		**Evalb**			
	Sent.	Corpus	LP	LR	F1	EX%
PAPCFG	0.857	0.840	73.5	72.8	73.1	14.5
SplitPCFG	0.870	0.853	77.9	77.1	77.5	16.0
Berkeley	**0.905**	**0.894**	**83.9**	83.4	**83.6**	24.0
DP-TSG	0.858	0.841	77.1	76.8	76.9	16.0
Stanford	0.869	0.853	78.5	79.6	79.0	17.6
Stanford+FactLex	0.877	0.860	79.0	79.6	79.3	19.6
Stanford+FactLex*	0.890	0.874	82.8	**84.0**	83.4	**27.4**

Table 10
Arabic MWE identification per category and overall results (test set, sentences ≤ 40 words).

	#gold	PAPCFG	SplitPCFG	Berkeley	DP-TSG	Stanford	FactLex	FactLex*
MWA	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MWP	34	36.9	76.9	81.2	**91.8**	88.2	88.2	86.6
MWN	465	9.8	66.7	74.6	**81.1**	76.6	77.0	77.5
Total:	500	13.2	67.4	74.8	**81.9**	77.5	77.9	78.2

non-MWE n-grams that occurred only once. For each resulting n-gram, we created real-valued feature vectors and trained a binary SVM classifier with Weka (Hall et al. 2009) with an RBF kernel. See Appendix D for further configuration details.

Results. Because our parsers mark MWEs as labeled spans, MWE identification is a by-product of parsing. Our evaluation metric is category-level Evalb for the MWE non-terminal categories. We report both the per-category scores (Tables 10 and 11), and a weighted average for all categories. Table 12 shows aggregate MWE identification results. All parsing models—even the baselines—exceed `mwetoolkit` by a wide margin.

7. Discussion

7.1 MWE Identification Results

The main contribution of this article is Table 12, which summarizes MWE identification results. For both languages, our parsing models yield substantial improvements over the n-gram classification method represented by `mwetoolkit`. The best improvements come from different models: The DP-TSG model achieves 66.9% F1 absolute improvement for Arabic and the Stanford+FactLex* achieves 50.0% F1 absolute improvement for French.

Differences in how the training resources were constructed may account for differences in the ordering of the models. The Arabic MWE list consists mainly of named entities and nominal compounds, hence the high concentration of MWN types in the

Table 11
French MWE identification per category and overall results (test set, sentences ≤ 40 words). MWI and MWCL do not occur in the test set.

	#gold	PAPCFG	SplitPCFG	Berkeley	DP-TSG	Stanford	FactLex	FactLex*
MWET	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MWV	26	6.1	56.1	54.3	56.2	57.1	44.9	**83.3**
MWA	8	**42.9**	29.6	36.7	36.0	26.1	25.0	33.3
MWN	457	41.1	56.0	67.4	65.7	64.8	64.9	**86.3**
MWD	15	60.0	70.3	**74.4**	65.1	68.4	64.9	70.3
MWPRO	17	83.9	70.3	**87.6**	75.3	72.2	72.2	81.3
MWADV	220	46.8	68.0	72.5	77.2	75.0	76.0	**87.9**
MWP	162	49.0	78.9	81.4	79.5	81.2	81.9	**92.9**
MWC	47	74.2	80.7	83.7	85.8	86.3	88.2	**97.9**
Total:	955	46.0	64.2	71.4	71.3	70.5	70.5	**87.3**

Table 12
MWE identification F1 of the parsing models vs. the `mwetoolkit` baseline (test set, sentences $\leq$ 40 words). FactLex* uses gold morphological analyses at test time.

Model	Arabic F1	French F1
`mwetoolkit` (baseline)	15.0	37.3
PAPCFG	13.2	46.0
SplitPCFG	67.4	64.2
Berkeley	74.8	71.4
DP-TSG	**81.9**	71.3
Stanford	77.5	70.5
Stanford+FactLex	77.8	70.5
Stanford+FactLex*	78.2	**87.3**

pre-processed ATB (Table 10). Consequently, this particular Arabic MWE identification experiment is similar to joint parsing and named entity recognition (NER) (Finkel and Manning 2009). The DP-TSG is effective at memorizing the entities and re-using them at test time. It would be instructive to compare the DP-TSG to the discriminative model of Finkel and Manning (2009), which currently represents the state-of-the-art for joint parsing and NER.

The Berkeley and DP-TSG models are equally effective at learning French MWE rules. One explanation for this result could be the CC tag set, which was explicitly tuned for the Berkeley parser. The CC tag set improved Berkeley MWE identification accuracy by 1.8% F1 and basic parsing accuracy by 1.2% F1 over the previous version of our work (Green et al. 2011), in which we used the basic FTB tag set. However, this tag set yielded only 0.2% F1 and 1.1% F1 improvements, respectively, for the DP-TSG.

Interpretation of the factored lexicon results should account for resource asymmetries. For French, the extraordinary result with gold analyses (Stanford+FactLex*) is partly due to annotation errors. Gold morphological analyses are missing for many of the MWE tokens in the FTB. The factored lexicon thus learns that when a token has no morphology, it is usually part of an MWE. In the automatic setting (Stanford+FactLex), however, Morfette tends to assign morphology to the MWE tokens because it has no semantic knowledge. Consequently, the morphological predictions are less consistent, and the parsing model falls back to the baseline Stanford result. Certainly more consistent FTB annotations would help Morfette, which we found to be significantly less accurate on our version of the FTB than MADA on the ATB (see Habash and Rambow 2005). Another remedy would be to incorporate MWE knowledge into the lexical analyzer, a strategy that Constant, Sigogne, and Watrin (2012) recently found to be very effective.

The Arabic factored lexicon results are more realistic. Stanford+FactLex* achieves a 0.7% F1 improvement over Stanford along with a significant improvement in exact match (EX%). In the automatic setting, a 0.3% F1 improvement is maintained for MWE identification. One direction for improvement might be the POS tag set. The "Kulick" tag set encodes some morphological information (e.g., number, definiteness), so the factored lexicon can be redundant. Eliminating this overlap might improve accuracy.

Table 13
Sample of human-interpretable Arabic TSG rules. Recursive rules like MWA→A MWA result from memoryless binarization of *n*-ary rules. This pre-processing step not only increases parsing accuracy, but also allows the generation of previously unseen MWEs of a given type.

MWN		MWP		MWA	
لس انجليس	'Los Angeles'	ب MWP	'with MWP'	رفيعة المستوى	'high-level'
رئيس الوزراء	'Prime Minister'	حتى الآن	'until now'	سوفياتية الصنع	'Soviet-made'
عسكري N	'military N'	بالتوقيت المحلي	'local time'	A MWA	
مجلس N الدولة	'national N council'	من ناحية اخرى	'on the other hand'		

7.2 Interpretability of DP-TSG MWE Rules

Arabic. Table 13 lists a sample of the TSG rules learned by the DP-TSG model. Fixed expressions such as names (*Los Angeles*) and titles (*Prime Minister*) are cached in the grammar. The model also generalizes over nominal compounds with rules like *military N*, which captures *military coup*, *military council*, and so forth. For multiword adjectives, the model caches several instances of false *iDafa* in full (*high-level*, *Soviet-made*). Memoryless binarization permits the grammar to capture rules like MWA → A MWA, which permits generation of a previously unseen multiword adjectives. Some of these recursive rules are lexicalized, as in the multiword preposition rule MWP → ب MWP.

French. We find that the DP-TSG model also learns useful generalizations over French MWEs. A sample of the rules is given in Table 14. Some specific sequences like "[$_{MWN}$ [coup de N]]" are part of the grammar: such rules can indeed generate quite a few MWEs, for example, *coup de pied* ('kick'), *coup de coeur*, *coup de foudre* ('love at first sight'), *coup de main* ('help'), *coup d'état*, *coup de grâce*. Certain of these MWEs are unseen in the training data. For MWV, the grammar contains "V de N" as in *avoir de cesse* ('give no peace'), *perdre de vue* [lose from sight] ('forget'), *prendre de vitesse* [take from speed] ('outpace'). For prepositions, the grammar stores full subtrees of MWPs, but also generalizes over very frequent sequences: "en N de" occurs in many multiword prepositions (e.g., *en compagnie de*, *en face de*, *en matière de*, *en terme de*, *en cours de*, *en faveur de*, *en raison de*, *en fonction de*). The TSG grammar thus provides a categorization of MWEs consistent with the Lexicon-Grammar. It also learns verbal phrases which contain discontiguous MWVs due to the insertion of an adverb or negation such as "[$_{VN}$ [$_{MWV}$ va] [$_{MWADV}$ d'ailleurs] [$_{MWV}$ bon train]]" [go indeed well], "[$_{VN}$ [$_{MWV}$ a] [$_{ADV}$ jamais] [$_{MWV}$ été question d']]" [has never been in question].

Table 14
Sample of human-interpretable French TSG rules.

MWN	MWV	MWP
sociétés de N	sous - V	de l'ordre de
chef de N	mis en N	y compris
coup de N	V DET N	au N de
N d'état	V de N	en N de
N de N	V en N	ADV de
N à N		

7.3 Basic Parsing Results

The relative rankings of the different models are the same for Arabic and French (Berkeley > Stanford parser > DP-TSG > PAPCFG), and these rankings correspond to those observed for English (Cohn, Blunsom, and Goldwater 2010). Although statistical statements cannot be made about the difficulty of parsing the two languages by comparing raw evaluation figures, we can compare the differences between PAPCFG and the best model for each language. From this perspective, manual rule splitting in the Stanford parser is apparently more effective for the ATB than for the FTB. Differences in annotation styles may account for this discrepancy. Consider the unbinarized treebanks. The ATB training set has 8,937 unique non-unary rule types with mean branching factor $M = 2.41$ and sample standard deviation $SD = 0.984$. The FTB has a flat annotation style, which leads to more rule types (16,159) with a higher branching factor ($M = 2.87$, $SD = 1.51$).

A high branching factor can lead to more brittle grammars, an empirical observation that motivated memoryless binarization in both the Berkeley parser (Petrov et al. 2006, page 434) and the DP-TSG. The Berkeley parser results also seem to support the observation that rule refinement is less effective for the FTB. Automatic rule refinement results in a 16.1% F1 absolute improvement over PAPCFG for Arabic, but only 10.0% F1 for French.

Of course, the FTB contains 28.5% fewer sentences than the ATB, so the FTB rule counts are also sparser. In addition, we found that the FTB has lower inner-annotator agreement (IAA) than the ATB (Appendix C), which also negatively affects supervised models. Finally, Evalb penalizes flat treebanks like the FTB (Rehbein and van Genabith 2007). To counteract that bias, we also included a Leaf Ancestor evaluation. Nonetheless, even Leaf Ancestor showed that, with respect to PAPCFG, the best Arabic model improved nearly twice as much as the best French model.

The DP-TSG improves over PAPCFG, but does not exceed the Berkeley parser. One crucial difference between the two models is the decoding objective. The Berkeley parser maximizes the expected rule count (max-rule-sum) (Petrov and Klein 2007), an objective that Cohn, Blunsom, and Goldwater (2010) demonstrated could improve the DP-TSG by 2.0% F1 over Viterbi for English *with no changes to the grammar*. We decoded with Viterbi, so our results are likely a lower bound relative to what could be achieved with objectives that correlate with labeled recall. Because MWE identification is a by-product of parsing, we expect that MWE identification accuracy would also improve.

Because the DP-TSG and PAPCFG have the same weak generative capacity, the improvement must come from relaxing independencies in the grammar rules (by saving larger tree fragments). This is the same justification for manual rule refinement for PCFGs (Johnson 1998, page 614). We observe an 8.5% F1 absolute improvement for Arabic, but just 3.8% F1 for French. Nonetheless, we chose this model precisely for its greater strong generative capacity, which we hypothesized would improve MWE identification accuracy. The MWE identification results seem to bear out this hypothesis.

8. Related Work

This section contains three parts. First, we review work on MWEs in linguistics and relate it to parallel developments in NLP. Second, we describe other syntax-based MWE identification methods. Finally, we enumerate related experiments on Arabic and French.

8.1 Analysis of MWEs in Linguistics and NLP

An underlying assumption of mainline generative grammatical theory is that words are the basic units of syntax (Chomsky 1957). Lexical insertion is the process by which words enter into phrase structure, thus lexical insertion rules have the form [N $\rightarrow$ *dog, car, apple*], and so on. This assumption, however, was questioned not long after it was proposed, as early work on idiomatic constructions like *kick the bucket*—which functions like a multiword verb in syntax—seemed to indicate a conflict (Katz and Postal 1963; Chafe 1968). Chomsky (1981) briefly engaged *kick the bucket* in a footnote, but idioms remained a peripheral issue in mainline generative theory.

To others, the marginal status of idioms and fixed expressions seemed inappropriate given their pervasiveness cross-linguistically. In their classic work on the English construction let alone, Fillmore, Kay, and O'Connor (1988) argued that the basic units of grammar are not Chomskyan rules but **constructions**, or triples of phonological, syntactic, and conceptual structures. The subsequent development of Construction Grammar (Fillmore, Kay, and O'Connor 1988; Goldberg 1995) maintained the central role of idioms. Jackendoff (1997) has advanced a linguistic theory, the Parallel Architecture, which includes multiword expressions in the lexicon.

In NLP, concurrent with the development of Construction Grammar, Scha (1990) conceptualized an alternate model of parsing in which new utterances are built from previously observed language fragments. In his model, which became known as **data-oriented parsing** (DOP) (Bod 1992), "idiomaticity is the rule rather than the exception" (Scha 1990, page 13). Most DOP work, however, has focused on parameter estimation issues with a view to improving overall parsing performance rather than explicit modeling of idioms.

Given developments in linguistics, and to a lesser degree DOP, in modeling MWEs, it is curious that most NLP work on MWE identification has not utilized syntax. Moreover, the *words-with-spaces* idea, which Sag et al. (2002) dismissed as unattractive on both theoretical and computational grounds,[22] has continued to appear in NLP evaluations such as dependency parsing (Nivre and Nilsson 2004), constituency parsing (Arun and Keller 2005), and shallow parsing (Korkontzelos and Manandhar 2010). In all cases, the conclusion was drawn that pre-grouping MWEs improves task accuracy. Because the yields (and thus the labelings) of the evaluation sentences were modified, however, the experiments were not strictly comparable. Moreover, gold pre-grouping was usually assumed, as was the case in most FTB parsing evaluations after Arun and Keller (2005).

The *words-with-spaces* strategy is especially unattractive for MRLs because (1) it intensifies the sparsity problem in the lexicon; and (2) it is not robust to morphological and syntactic processes such as inflection and phrasal expansion.

8.2 Syntactic Methods for MWE Identification

There is a voluminous literature on MWE identification, so we focus on syntax-based methods. The classic statistical approach to MWE identification, Xtract (Smadja 1993), used an incremental parser in the third stage of its pipeline to identify predicate-argument relationships. Lin (1999) applied information-theoretic measures to automatically extracted dependency relationships to find MWEs. To our knowledge,

22 Sag et al. (2002) showed how to integrate MWE information into a non-probabilistic head-driven phrase structure grammar for English.

Wehrli (2000) was the first to propose the use of a syntactic parser for multiword expression identification. No empirical results were provided, however, and the MWE-augmented *scoring function* for the output of his symbolic parser was left to future research. Recently, Seretan (2011) used a symbolic parser for collocation extraction. Collocations are two-word MWEs. In contrast, our models handle arbitrary length MWEs.

To our knowledge, only two previous studies considered MWEs in the context of statistical parsing. Nivre and Nilsson (2004) converted a Swedish corpus into two versions: one in which MWEs were left as tokens, and one in which they were grouped (*words-with-spaces*). They parsed both versions with a transition-based parser, showing that the words-with-spaces version gave an improvement over the baseline. Cafferkey (2008) also investigated the words-with-spaces idea along with imposing chart constraints for pre-bracketed spans. He annotated the PTB using external MWE lists and an NER system, but his technique did not improve two different constituency models. At issue in both of these studies is the comparison to the baseline. MWE pre-grouping changes the number of evaluation units (dependency arcs or bracketed spans), thus the results are not strictly comparable. From an application perspective, pre-grouping assumes high accuracy identification, which may not be available for all languages.

Our goal differs considerably from these two studies, which attempt to improve parsing via MWE information. In contrast, we tune statistical parsers for MWE identification.

8.3 Related Experiments on Arabic and French

Arabic Statistical Constituency Parsing. Kulick, Gabbard, and Marcus (2006) were the first to parse the sections of the ATB used in this article. They adapted the Bikel parser (Bikel 2004) and improved accuracy primarily through punctuation equivalence classing and the Kulick tag set. The ATB was subsequently revised (Maamouri, Bies, and Kulick 2008), and Maamouri, Bies, and Kulick (2009) produced the first results on the revision for our split of the revised corpus. They only reported development set results with gold POS tags, however. Petrov (2009) adapted the Berkeley parser to the ATB, and we later provided a parameterization that dramatically improved his baseline (Green and Manning 2010). We also adapted the Stanford parser to the ATB, and provided the first results for non-gold tokenization. Attia et al. (2010a) developed an Arabic unknown word model for the Berkeley parser based on signatures, much like those in Table 3. More recently, Huang and Harper (2011) presented a discriminative lexical model for Arabic that can encode arbitrary local lexical features.

Arabic MWE Identification. Very little prior work exists on Arabic MWE identification. Attia (2006) demonstrated a method for integrating MWE knowledge into a lexical-functional grammar, but gave no experimental results. Siham Boulaknadel and Aboutajdine (2008) evaluated several lexical association measures in isolation for MWE identification in newswire. More recently, Attia et al. (2010b) compared cross-lingual projection methods (using Wikipedia and English Wordnet) with standard n-gram classification methods.

French Statistical Constituency Parsing. Abeillé (1988) and Abeillé and Schabes (1989) identified the linguistic and computational attractiveness of lexicalized grammars for modeling non-compositional constructions in French well before DOP. They developed a small tree adjoining grammar (TAG) of 1,200 elementary trees and 4,000 lexical items that included MWEs. Recent statistical parsing work on French has included stochastic tree insertion grammars (STIG), which are related to TAGs, but with a restricted adjunction operation.[23] Both Seddah, Candito, and Crabbé (2009) and Seddah (2010) showed that STIGs underperform CFG-based parsers on the FTB. In their experiments, MWEs were grouped. Appendix B describes additional prior work on CFG-based FTB parsing.

French MWE Identification. Statistical French MWE identification has only been investigated recently. We previously reported the first results on the FTB using a parser for MWE identification (Green et al. 2011). Contemporaneously, Watrin and Francois (2011) applied *n*-gram methods to a French corpus of multiword adverbs (Laporte, Nakamura, and Voyatzi 2008). Constant and Tellier (2012) used a linear chain conditional random fields model (CRF) for joint POS tagging and MWE identification. They incorporated external linguistic resources as features, but reported results for a much older version of the FTB. Subsequently, Constant, Sigogne, and Watrin (2012) integrated the CRF model into the Berkeley parser and evaluated on the pre-processed FTB used in this article. Their best model (with external lexicon features) achieved 77.8% F1.

9. Conclusion

In this article, we showed that parsing models are very effective for identifying arbitrary-length, contiguous MWEs. We achieved a 66.9% F1 absolute improvement for Arabic, and a 50.0% F1 absolute improvement for French over *n*-gram classification methods. All parsing models discussed in the paper improve MWE identification over *n*-gram methods, but the best improvements come from different models. Unlike *n*-gram classification methods, parsers provide syntactic subcategorization and do not require heuristic pre-filtering of the training data. Our techniques can be applied to any language for which the following linguistic resources exist: a syntactic treebank, an MWE list, and a morphological analyzer.

More fundamentally, we exploited a connection between syntax and idiomatic semantics. This connection has been debated in linguistics, yet overlooked in statistical NLP until now. Although empirical task evaluations do not always reinforce linguistic theories, our results suggest that syntactic context can help identify idiomatic language, as posited by some modern grammar theories.

We introduced the factored lexicon for the Stanford parser, a simple extension to the lexical insertion model that helps combat lexical sparsity in morphologically rich languages. In the gold setting, the factored lexicon yielded improvements over the basic lexicon for both standard parsing and MWE identification. Results were lower in the automatic setting, suggesting that it might be helpful to optimize the morphological analyzers for specific features included in downstream tasks like parsing. We evaluated on in-domain data, but we expect that the factored lexicon would be even more useful on out-of-domain text with higher out-of-vocabulary rates.

We have also provided empirical evidence that TSGs can capture idiomatic usage as well as or better than a state-of-the-art CFG-based parser. The suitability of TSGs for idioms has been discussed since the earliest days of DOP (Scha 1990), but it has never been demonstrated with experiments like ours. Although the DP-TSG, which is a relatively new parsing model, still lags other parsers in terms of overall labeling

23 Unlike TAG and TIG, TSG does not include an adjunction operator.

accuracy, we have shown that it is already very effective for tasks like MWE identification. Because we modified the syntactic representation rather than the model formulation, general improvements to this parsing model should yield improvements in MWE identification accuracy.

Appendix A: Additional French MWEs

This appendix describes the other French MWE categories annotated in the FTB.

Adverbial idioms (MWADV) often start with a preposition (Example (11)) but can have very different part-of-speech sequences:

(11) a. P N: *du coup* ('so'), *sans doute* ('doubtless')
b. P D A N: *avec un bel ensemble* [with a nice ensemble] ('in harmony')
c. P ADV P ADV: *de plus en plus* ('more and more')
d. V V: *peut-être* [can be] ('maybe')
e. ADV A: *bien sûr* [very certain] ('of course')
f. ET ET: *a priori*, *grosso modo*

Foreign words (MWET) include English nominal idioms, such as *cash flow* and *success story*, which are less integrated in French than words such as *T-shirt*. Expressions such as *Just do it* or *struggle for life* also fall in this category.

Prepositional idioms (MWP) are mostly fixed (Example (12)), but some permit minimal variation such as *de* vs. *des* or *à* vs. *au*:

(12) a. P N P: *en dépit de* ('despite'), *à hauteur de* ('at the height of')
b. P D N P: *dans le cadre de* ('in the framework of'), *à l'exception de* ('at the exception of')
c. P P: *afin de* ('to'), *jusqu'à* ('as far as')
d. ADV P: *autour de* ('around'), *quant à* ('as for')
e. N V P: *compte tenu de* ('taking into account'), *exception faite de* [exception made of] ('at the exception of')

Pronominal idioms (MWPRO) consist of demonstrative pronouns (*celui-ci* 'this one', *celui-là* 'that one') and reflexive pronouns (*lui-même* 'himself'), which vary in gender and number, as well as a few indefinite pronouns which allow gender inflection (*d'aucuns* 'no-one', *quelque chose* 'something', *qui que ce soit* 'who ever it is', *n'importe qui* 'anyone') and some which are fixed (*d'autres* 'others', *la plupart* 'most', *tout un chacun* 'everyone', *tout le monde* 'everybody').

Multiword determiners (MWD) consist of expressions such as *bien des* ('a lot of') and *tout le* ('all the'), which display minimal variation in terms of inflection (e.g., *la plupart de* vs. *la plupart des* 'most of'). Numbers which act as determiners in the sentence (*classées en vingt-huit catégories* 'categorized in twenty-eight categories') are also classified as MWD.

Multiword conjunctions (MWC) are a fixed class:

(13) a. C C: *parce que* ('because')
b. ADV C: *même si* ('even so')
c. V C: *pourvu que* ('so long as')
d. D N C: *au moment où* ('at the time when')
e. CL V A C: *il est vrai que* ('it's true that')
f. ADV C ADV ADV C: *tant et si bien que* ('to such an extent that')

Multiword interjections (MWI) are a small category with expressions such as *mille sabords* ('blistering barnacles') and *au secours* ('help').

Appendix B: Comparison to Prior FTB Pre-Processing

Our FTB pre-processing is automatic, unlike all previous methods.

ARUN-CONT and *ARUN-EXP*. (Arun and Keller 2005) Two versions of the full 20,000-sentence treebank that differed principally in their treatment of MWEs: (1) CONT, in which MWE tokens were grouped (*en moyenne* → *en_moyenne*); and (2) EXP, in which MWEs were marked with a flat structure. For both representations, they also gave results in which coordinated phrase structures were flattened. In the published experiments, they mistakenly removed half of the corpus, believing that the multi-terminal (per POS tag) annotations of MWEs were XML errors (Schluter and van Genabith 2007).

MFT. (Schluter and van Genabith 2007) Manual revision to 3,800 sentences. Major changes included coordination raising, an expanded POS tag set, and the correction of annotation errors. Like ARUN-CONT, MFT contains concatenated MWEs.

FTB-UC. (Candito and Crabbé 2009) A version of the functionally annotated section that makes a distinction between MWEs that are "syntactically regular" and those that are not. Syntactically regular MWEs were given internal structure, whereas all other MWEs were grouped. For example, nouns followed by adjectives, such as *loi agraire* ('land law') or *Union monétaire et économique* ('monetary and economic Union') were considered syntactically regular. They are MWEs because the choice of adjective is arbitrary (*loi agraire* and not **loi agricole*, similarly to ('coal black') but not (*'crow black'), for example), but their syntactic structure is not intrinsic to MWEs. In such cases, FTB-UC gives the MWE a conventional analysis of an NP with internal structure. Such analysis is indeed sufficient to recover the meaning of these semantically compositional MWEs that are extremely productive. FTB-UC loses information about MWEs with non-compositional semantics, however.

Almost all work on the FTB has followed ARUN-CONT and used gold MWE pre-grouping. Candito, Crabbé, and Denis (2010) were the first to acknowledge and address this issue, but they still used FTB-UC (with some pre-grouped MWEs). Because the syntax and definition of MWEs is a contentious issue, we take a more agnostic view—which is consistent with that of the FTB annotators—and leave them ungrouped. This permits a data-oriented approach to MWE identification that is more robust to changes to the status of specific MWE instances.

Although our FTB basic parsing results are lower than those of Seddah (2010), the experiments are not comparable: The data split and pre-processing were different, and he grouped MWEs.

Appendix C: Annotation Consistency of Treebanks

Differences in annotation quality among corpora complicate cross-lingual experimental comparisons. To control for this variable, we performed an annotation consistency evaluation on the PTB, ATB, and FTB. The conventional wisdom has it that the PTB has comparatively high inter-annotator agreement (IAA). In the initial release of the ATB, IAA was inferior to other LDC treebanks, although in subsequent revisions, IAA was quantifiably improved (Maamouri, Bies, and Kulick 2008). The FTB also had significant annotation errors upon release (Arun and Keller 2005), but it, too, has been revised.

To quantify IAA, we extend the **variation nucleus** method of Dickinson (2005) to compare annotation error rates. Let $\mathcal{C}$ be a set of tuples $\langle s, l, i \rangle$, where s is a substring at corpus position i with label l. We consider all substrings in the corpus. If s is bracketed at position i, then its label is its non-terminal category. Otherwise, s has label $l = \texttt{NIL}$. To locate variation nuclei, define L_s as the set of all labels associated with each unique s. If $|L_s| > 1$, then s is a variation nucleus.[24]

Variation nuclei can result from either annotation errors or linguistic ambiguity. Human evaluation is one way to distinguish between the two cases. Following Dickinson (2005), we sampled 100 variation nuclei from each corpus and evaluated each sample for the presence of an annotation error. To control for the number of corpus positions included in each treebank sample, we used frequency-matched stratified sampling with bin sizes of 2, 3, 4, 10, 50, and 500.

The human evaluators were a non-native, fluent Arabic speaker for the ATB (the first author), a native French speaker for the FTB (the second author), and a native English speaker for the WSJ (the third author).[25] Table C.1 shows the results of the evaluation, which supports the anecdotal consistency ranking of the three treebanks.[26] The FTB averages more than one variation nucleus per sentence and has twice the token-level error rate of the other two treebanks.

Appendix D: `mwetoolkit` Configuration

We configured `mwetoolkit`[27] with the four standard lexical features: the maximum likelihood estimator, Dice's coefficient, pointwise mutual information, and Student's *t*-score. We added the POS sequence for each *n*-gram as a single feature. We removed the Web counts features since the parsers do not use auxiliary data.

Because MWE *n*-grams only account for a small fraction of the *n*-grams in the corpus, we filtered the training and test sets by removing all *n*-grams that occurred once. To further balance the proportion of MWEs, we trained on all valid MWEs plus 10x randomly selected non-MWE *n*-grams. This proportion matches the fraction

24 Kulick, Bies, and Mott (2011) extended our method with TAGs to account for nested bracketing errors.

25 Unlike Dickinson (2005), we stripped traces and only considered POS tags when pre-terminals were the only intervening nodes between the nucleus and its bracketing (e.g., unaries, base NPs). Because our objective was to compare distributions of bracketing discrepancies, we did not prune the set of nuclei.

26 The total variation nuclei in each corpus were: 22,521 (WSJ), 15,629 (ATB), and 14,803 (FTB).

27 We re-implemented `mwetoolkit` in Java for compatibility with Weka and our pre-processing routines.

Table C.1
Evaluation of 100 randomly sampled variation nuclei for training splits of the WSJ, ATB, and FTB. **Corpus positions** indicates the number of corpus positions in the sample (a variation nucleus by definition appears in at least two corpus positions). **Nuclei per tree** is the average nuclei per syntactic tree in the corpus, a statistic that gives a rough estimate of variability across the corpus. The **type-level** error rate indicates the number of variation nuclei for which at least one error existed. The **token-level** error rate indicates the ratio of errors to corpus positions. We computed 95% confidence intervals for the type-level error rate.

	Corpus Positions	Nuclei Per Tree	Error % Type	Error % Token	Type 95% Confidence Interval
PTB (2-21)	750	0.565	16.0%	4.10%	[8.80%, 23.2%]
ATB (train)	658	0.830	26.0%	4.00%	[17.4%, 34.6%]
FTB (train)	668	1.10	28.0%	9.13%	[19.2%, 36.8%]

of MWE/non-MWE tokens in the FTB. As we generated a random training set, we reported the average of three independent training runs.

We created feature vectors for the training *n*-grams and trained a binary SVM classifier with Weka (Hall et al. 2009). Although `mwetoolkit` defaults to a linear kernel, we achieved higher accuracy on the development set with an RBF kernel.

The FTB is sufficiently large for the corpus-based methods implemented in `mwetoolkit`. Ramisch, Villavicencio, and Boitet (2010) experimented with the Genia corpus, which contains 18k English sentences and 490k tokens, similar to the FTB. Their test set had 895 sentences, fewer than ours. They reported 30.6% F1 for their task against an Xtract baseline, which only obtained 7.3% F1. Their best result compares favorably (in magnitude) to our `mwetoolkit` baselines for French and Arabic.

Acknowledgments

We thank John Bauer for material contributions to the MWE identification experiments, and Claude Reichard for help with editing this article. We also thank Marie Candito, Markus Dickinson, Chris Dyer, Ali Farghaly, Dan Flickinger, Nizar Habash, Seth Kulick, Beth Levin, Percy Liang, David McClosky, Carlos Ramisch, Ryan Roth, Djamé Seddah, Valentin Spitkovsky, and Reut Tsarfaty for insightful comments on previous versions of this work. The first author was supported by a National Science Foundation Graduate Research Fellowship. The second author was supported by a Stanford Interdisciplinary Graduate Fellowship.

References

Abeillé, A. 1988. Parsing French with Tree Adjoining Grammar: some linguistic accounts. In *COLING*, pages 7–12, Budapest, Hungary.

Abeillé, A., L. Clément, and A. Kinyon. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks: building and using parsed corpora*. Kluwer, chapter 10.

Abeillé, A. and Y. Schabes. 1989. Parsing idioms in lexicalized TAGs. In *EACL*, pages 1–9, Manchester.

Arun, A. 2004. Statistical parsing of the French treebank. Master's thesis, University of Edinburgh.

Arun, A. and F. Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *ACL*, pages 306–313, Ann Arbor, MI.

Ashraf, A. 2012. *Arabic Idioms: A Corpus-Based Study*. Routledge.

Attia, M. 2006. Accommodating multiword expressions in an Arabic LFG grammar. In *Advances in Natural Language Processing*, volume 4139. Springer, pages 87–98.

Attia, M., J. Foster, D. Hogan, J. Le Roux, L. Tounsi, and J. van Genabith. 2010a. Handling unknown words in statistical latent-variable parsing models for Arabic, English and French. In *First Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL)*, pages 67–75, Los Angeles, CA.

Attia, M., A. Toral, L. Tounsi, P. Pecina, and J. van Genabith. 2010b. Automatic extraction of Arabic multiword expressions. In *Workshop on Multiword Expressions: From Theory to Applications*, pages 19–27, Beijing.

Baldwin, T. and S. N. Kim. 2010. Multiword expressions. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press, chapter 12, pages 267–293.

Bansal, M. and D. Klein. 2010. Simple, accurate parsing with an all-fragments grammar. In *ACL*, pages 1098–1107, Uppsala.

Bikel, D. M. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4):479–511.

Bilmes, J. and K. Kirchoff. 2003. Factored language models and generalized parallel backoff. In *NAACL*, pages 4–6, Edmonton.

Blunsom, P. and T. Baldwin. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *EMNLP*, pages 164–171, Sydney.

Bod, R. 1992. A computation model of language performance: Data-Oriented Parsing. In *COLING*, pages 855–859, Nantes.

Cafferkey, C. 2008. Exploiting multi-word units in statistical parsing and generation. Master's thesis, Dublin City University.

Candito, M. and B. Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *IWPT*, pages 138–141, Paris.

Candito, M., B. Crabbé, and P. Denis. 2010. Statistical French dependency parsing: Treebank conversion and first results. In *LREC*, pages 1840–1847, Valletta.

Candito, M. and D. Seddah. 2010. Parsing word clusters. In *First Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL)*, pages 76–84, Los Angeles, CA.

Carpuat, M. and M. Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *HLT-NAACL*, pages 242–245, Los Angeles, CA.

Chafe, W. L. 1968. Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language*, 4(2):109–127.

Chomsky, N. 1957. *Syntactic Structures*. Mouton, London.

Chomsky, N. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Foris Publications, Holland.

Chrupala, G., G. Dinu, and J. van Genabith. 2008. Learning morphology with Morfette. In *LREC*, pages 2362–2367, Marrakech.

Cohn, T., P. Blunsom, and S. Goldwater. 2010. Inducing tree-substitution grammars. *JMLR*, 11:3053–3096.

Cohn, T., S. Goldwater, and P. Blunsom. 2009. Inducing compact but accurate tree-substitution grammars. In *HLT-NAACL*, pages 548–556, Boulder, CO.

Constant, M., A. Sigogne, and P. Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *ACL*, pages 204–212, Jeju.

Constant, M. and I. Tellier. 2012. Evaluating the impact of external lexical resources into a CRF-based multiword segmenter and part-of-speech tagger. In *LREC*, pages 646–650, Istanbul.

Crabbé, B. and M. Candito. 2008. Expériences d'analyse syntaxique statistique du français. In *TALN*, pages 1–10, Avignon.

Dickinson, M. 2005. *Error Detection and Correction in Annotated Corpora*. Ph.D. thesis, The Ohio State University.

Dybro-Johansen, A. 2004. Extraction automatique de grammaires à partir d'un corpus français. Master's thesis, Université Paris 7.

Dyer, C., A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik. 2010. `cdec`: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *ACL System Demonstrations*, pages 7–12, Uppsala.

Evert, S. 2008. The MWE 2008 Shared Task: Ranking MWE candidates. In *Presentation at 2008 Workshop on Multiword Expressions*, Marrakech.

Fillmore, C. J., P. Kay, and M. C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, 64(3):501–538.

Finkel, J. R. and C. D. Manning. 2009. Joint parsing and named entity recognition. In *HLT-NAACL*, pages 326–334, Boulder, CO.

Goldberg, A. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University Of Chicago Press, Chicago.

Green, S., M-C. de Marneffe, J. Bauer, and C. D. Manning. 2011. Multiword expression identification with Tree Substitution Grammars: A parsing *tour de force* with French. In *EMNLP*, pages 725–735, Edinburgh.

Green, S. and C. D. Manning. 2010. Better Arabic parsing: Baselines, evaluations,

and analysis. In *COLING*, pages 394–402, Beijing.
Gross, M. 1984. Lexicon-Grammar and the syntactic analysis of French. In *COLING-ACL*, pages 275–282, Stanford, CA.
Gross, M. 1986. Lexicon-Grammar: The representation of compound words. In *COLING*, pages 1–6, Bonn.
Habash, N. and O. Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *ACL*, pages 573–580, Ann Arbor, MI.
Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter*, 11:10–18.
Hogan, D., C. Cafferkey, A. Cahill, and J. van Genabith. 2007. Exploiting multi-word units in history-based probabilistic generation. In *EMNLP-CoNLL*, pages 267–276, Prague.
Huang, Z. and M. Harper. 2011. Feature-rich log-linear lexical model for latent variable PCFG grammars. In *IJCNLP*, pages 219–227, Chiang Mai.
Jackendoff, R. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.
Johnson, M. 1998. PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.
Katz, J. J. and P. M. Postal. 1963. Semantic interpretation of idioms and sentences containing them. *M.I.T. Research Laboratory of Electronics Quarterly Progress Report*, 70:275–282.
Klein, D. and C. D. Manning. 2003. Accurate unlexicalized parsing. In *ACL*, pages 423–430, Sapporo.
Koehn, P. and H. Hoang. 2007. Factored translation models. In *EMNLP-CoNLL*, pages 868–876, Prague.
Korkontzelos, I. and S. Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *HLT-NAACL*, pages 636–644, Los Angeles, CA.
Kübler, S. 2005. How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges. In *RANLP*, pages 79–88, Borovets.
Kulick, S., A. Bies, and J. Mott. 2011. Using derivation trees for treebank error detection. In *ACL*, pages 693–698, Portland, OR.
Kulick, S., R. Gabbard, and M. Marcus. 2006. Parsing the Arabic Treebank: Analysis and improvements. In *TLT*, pages 31–42, Prague.
Laporte, E., T. Nakamura, and S. Voyatzi. 2008. A French corpus annotated for multiword expressions with adverbial function. In *LREC Linguistic Annotation Workshop*, pages 48–51, Marrakech.
Levy, R. and G. Andrew. 2006. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *LREC*, pages 2,231–2,234, Genoa.
Levy, R. and C. D. Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *ACL*, pages 439–446, Sapporo.
Liang, P., M. I. Jordan, and D. Klein. 2010. Type-based MCMC. In *HLT-NAACL*, pages 573–581, Los Angeles, CA.
Lin, D. 1999. Automatic identification of non-compositional phrases. In *ACL*, pages 317–324, College Park, MD.
Maamouri, M., A. Bies, T. Buckwalter, and W. Mekki. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR*, pages 1–8, Cairo.
Maamouri, M., A. Bies, and S. Kulick. 2008. Enhancing the Arabic Treebank: A collaborative effort toward new annotation guidelines. In *LREC*, pages 3,192–3,196, Marrakech.
Maamouri, M., A. Bies, and S. Kulick. 2009. Creating a methodology for large-scale correction of treebank annotation: The case of the Arabic Treebank. In *MEDAR*, pages 138–144, Cairo.
Marantz, A. 1997. No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. In *21st Annual Penn Linguistics Colloquium*, pages 1–15, Philadelphia, PA.
Marcus, M., M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.
Nivre, J. and J. Nilsson. 2004. Multiword units in syntactic parsing. In *Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*, pages 1–8, Lisbon.
O'Donnell, T. J., J. B. Tenenbaum, and N. D. Goodman. 2009. Fragment grammars: Exploring computation and reuse in language. Technical report, MIT Computer Science and Artificial Intelligence Laboratory Technical Report Series, MIT-CSAIL-TR-2009-013.

O'Grady, W. 1998. The syntax of idioms. *Natural Language and Linguistic Theory*, 16:279–312.

Pecina, P. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

Petrov, S. 2009. *Coarse-to-Fine Natural Language Processing*. Ph.D. thesis, University of California-Berkeley.

Petrov, S., L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *ACL*, pages 443–440, Sydney.

Petrov, S. and D. Klein. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*, pages 404–411, Rochester, MN.

Post, M. and D. Gildea. 2009. Bayesian learning of a tree substitution grammar. In *ACL-IJCNLP, Short Papers*, pages 45–48, Suntec.

Rambow, O., D. Chiang, M. Diab, N. Habash, R. Hwa, K. Sima'an, V. Lacey, R. Levy, C. Nichols, and S. Shareef. 2005. Parsing Arabic dialects. Technical report. Johns Hopkins University.

Ramisch, C., A. Villavicencio, and C. Boitet. 2010. `mwetoolkit`: A framework for multiword expression identification. In *LREC*, pages 662–669, Valletta.

Rehbein, I. and J. van Genabith. 2007. Treebank annotation schemes and parser evaluation for German. In *EMNLP-CoNLL*, pages 630–639, Prague.

Ryding, K. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press.

Sag, I. A., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *CICLing*, pages 1–15, Mexico City.

Sampson, G. and A. Babarczy. 2003. A test of the leaf-ancestor metric for parse accuracy. *Natural Language Engineering*, 9:365–380.

Scha, R. 1990. Taaltheorie en taaltechnologie: competence en performance. In Q. A. M. de Kort and G. L. J. Leerdam, editors, *Computertoepassingen in de Neerlandistiek*. Landelijke Vereniging van Neerlandici (LVVNjaarboek), pages 7–22.

Schluter, N. and J. van Genabith. 2007. Preparing, restructuring, and augmenting a French treebank: Lexicalised parsers or coherent treebanks? In *Pacling*, pages 1–10, Melbourne.

Seddah, D. 2010. Exploring the Spinal-STIG model for parsing French. In *LREC*, pages 1,936–1,943, Valletta.

Seddah, D., M. Candito, and B. Crabbé. 2009. Cross parser evaluation and tagset variation: a French treebank study. In *IWPT*, pages 150–161, Paris.

Seddah, D., G. Chrupała, Ö. Çetinoglu, J. Genabith, and M. Candito. 2010. Lemmatization and lexicalized statistical parsing of morphologically rich languages: The case of French. In *First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL)*, pages 85–93, Los Angeles, CA.

Seretan, V. 2011. *Syntax-Based Collocation Extraction*. Springer.

Siham Boulaknadel, B. D. and D. Aboutajdine. 2008. A multi-word term extraction program for Arabic language. In *LREC*, pages 1,485–1,488, Marrakech.

Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*, pages 173–180, Edmonton.

Vijay-Shanker, K. and D. J. Weir. 1993. The use of shared forests in tree adjoining grammar parsing. In *EACL*, pages 384–393, Utrecht.

Watrin, P. and T. Francois. 2011. An *n*-gram frequency database reference to handle MWE extraction in NLP applications. In *Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 83–91, Portland, OR.

Wehrli, E. 2000. Parsing and collocations. In *Natural Language Processing–NLP 2000*, volume 1835 of *Lecture Notes in Computer Science*. Springer, pages 272–282.

West, M. 1995. Hyperparameter estimation in Dirichlet process mixture models. Technical report. Duke University.