Capitalization Cues Improve Dependency Grammar Induction

Valentin I. Spitkovsky
Stanford University and Google Inc.
valentin@cs.stanford.edu

Hiyan Alshawi
Google Inc., Mountain View, CA, 94043
hiyan@google.com

Daniel Jurafsky
Stanford University, Stanford, CA, 94305
jurafsky@stanford.edu

Abstract

We show that orthographic cues can be helpful for unsupervised parsing. In the Penn Treebank, transitions between upper- and lower-case tokens tend to align with the boundaries of base (English) noun phrases. Such signals can be used as partial bracketing constraints to train a grammar inducer: in our experiments, directed dependency accuracy increased by 2.2% (average over 14 languages having case information). Combining capitalization with punctuation-induced constraints in inference further improved parsing performance, attaining state-of-the-art levels for many languages.

1 Introduction

Dependency grammar induction and related problems of unsupervised syntactic structure discovery are attracting increasing attention (Rasooli and Faili, 2012; Mareček and Zabokrtský, 2011, inter alia). Since sentence structure is underdetermined by raw text, there have been efforts to simplify the task, via (i) pooling features of syntax across languages (Cohen et al., 2011; McDonald et al., 2011; Cohen and Smith, 2009); as well as (ii) identifying universal rules (Naseem et al., 2010) — such as verbo-centricity (Gimpel and Smith, 2011) — that need not be learned at all. Unfortunately most of these techniques do not apply to plain text, because they require knowing, for example, which words are verbs.

As standard practice shifts away from relying on gold part-of-speech (POS) tags (Seginer, 2007; Ponvert et al., 2010; Søgaard, 2011b; Spitkovsky et al., 2011c, inter alia), lighter cues to inducing linguistic structure become more important. Examples of useful POS-agnostic clues include punctuation boundaries (Ponvert et al., 2011; Spitkovsky et al., 2011b; Briscoe, 1994) and various kinds of bracketing constraints (Naseem and Barzilay, 2011; Spitkovsky et al., 2010b; Pereira and Schabes, 1992). We propose adding capitalization to this growing list of sources of partial bracketings. Our intuition stems from English, where (maximal) spans of capitalized words — such as Apple II, World War I, Mayor William H. Hudnut III, International Business Machines Corp. and Alexandria, Va — tend to demarcate proper nouns.

Consider a motivating example (all of our examples are from WSJ) without punctuation, in which all (eight) capitalized word clumps and uncased numerals match base noun phrase constituent boundaries:

\[
\text{NP}\text{Jay Stevens} \quad \text{of} \quad \text{NP}\text{Dean Witter} \quad \text{actually cut his per-share earnings estimate to} \quad \text{NP}\$9 \quad \text{from} \quad \text{NP}\$9.50 \quad \text{for} \quad \text{NP}\text{1989} \quad \text{and to} \quad \text{NP}\$9.50 \quad \text{from} \quad \text{NP}\$10.35 \quad \text{in} \quad \text{NP}\text{1990} \quad \text{because he decided sales would be even weaker than he had expected.}
\]

and another (whose first word happens to be a leaf), where capitalization complements punctuation cues:

\[
\text{NP}\text{Jurors} \quad \text{in} \quad \text{NP}\text{U.S. District Court} \quad \text{in} \quad \text{NP}\text{Miami} \quad \text{cleared} \quad \text{NP}\text{Harold Hershshenson}, \quad \text{a former executive vice president;} \quad \text{NP}\text{John Pagones}, \quad \text{a former vice president;} \quad \text{and} \quad \text{NP}\text{Stephen Vadas} \quad \text{and} \quad \text{NP}\text{Dean Ciporkin}, \quad \text{who had been engineers with} \quad \text{NP}\text{Cordis}.
\]

Could such chunks help bootstrap grammar induction and/or improve the accuracy of already-trained unsupervised parsers? In answering these questions, we will focus predominantly on sentence-internal capitalization. But we will also show that first words — those capitalized by convention — and uncased segments — whose characters are not even drawn from an alphabet — could play a useful role as well.

2 English Capitalization from a Treebank

We began our study by consulting the 51,558 parsed sentences of the WSJ corpus (Marcus et al., 1993): 30,691 (59.5%) of them contain non-trivially capitalized fragments — maximal (non-empty and not
sentence-initial) consecutive sequences of words that each differs from its own lower-cased form. Nearly all — 59,388 (96.2%) — of the 61,731 fragments are dominated by noun phrases; slightly less than half — 27,005 (43.8%) — perfectly align with constituent boundaries in the treebank; and about as many — 27,230 (44.1%) are multi-token. Table 1 shows the top POS sequences comprising fragments.

### 3 Analytical Experiments with Gold Trees

We gauged the suitability of capitalization-induced fragments for guiding dependency grammar induction by assessing accuracy, in WSJ,\(^1\) of parsing constraints derived from their end-points. Following the suite of increasingly-restrictive constraints on how dependencies may interact with fragments, introduced by Spitkovsky et al. (2011b, §2.2), we tested several such heuristics. The most lenient constraint, thread, only asks that no dependency path from the root to a leaf enter the fragment twice; tear requires any incoming arcs to come from the same side of the fragment; sprawl demands that there be exactly one incoming arc; loose further constrains any outgoing arcs to be from the fragment’s head; and strict — the most stringent constraint — bans external dependents. Since only strict is binding for single words, we experimented also with strict’: applying strict solely to multi-token fragments (ignoring singletons). In sum, we explored six ways in which dependency parse trees can be constrained by fragments whose end-points could be defined by capitalization (or in other various ways, e.g., semantic an-

---

1\(^{\text{We converted labeled constituents into unlabeled dependencies using deterministic “head-percolation” rules (Collins, 1999), discarding any empty nodes, etc., as is standard practice.}}\)

---

Table 1: Top 10 fragments of POS tag sequences in WSJ.

| Count | POS Sequence | Frac | Cum |
|-------|--------------|------|-----|
| 1     | 27,524       | NNP  | 44.6%|
| 2     | 17,222       | NNP NNP | 72.5 |
| 3     | 4,598        | NNP NNP NNP | 7.9 |
| 4     | 2,973        | JJ   | 84.8 |
| 5     | 1,716        | NNP NNP NNP NNP | 87.5 |
| 6     | 1,037        | NN   | 89.2 |
| 7     | 932          | PRP  | 90.7 |
| 8     | 846          | NNP  | 92.1 |
| 9     | 604          | NNP NNP | 93.1 |
| 10    | 526          | NNP NNP NNP NNP NNP | 93.9 |

WSJ +3,753 more with Count ≤ 408 6.1%

---

Table 2: Several sources of fragments’ end-points and %-correctness of their derived constraints (for English).

| Fragments | markup | punctl. | capital | initial | uncased |
|-----------|--------|---------|---------|---------|---------|
| thread    | 98.5   | 95.0    | 99.5    | 98.4    | 99.2    |
| tear      | 97.9   | 94.7    | 98.6    | 98.4    | 98.5    |
| sprawl    | 95.1   | 92.9    | 98.2    | 97.9    | 96.4    |
| loose     | 87.5   | 74.0    | 97.9    | 96.9    | 96.4    |
| strict    | 32.7   | 35.6    | 38.7    | 40.3    | 55.6    |
| strict’   | 35.6   | 39.2    | 59.3    | 66.9    | 61.1    |

---

Table 3: Supervised (directed) accuracy on Section 23 of WSJ using capitalization-induced constraints (vertical) jointly with punctuation (horizontal) in Viterbi-decoding.

\[\begin{array}{|c|c|c|c|c|}
\hline
\text{punct.}: \text{thread} & \text{tear} & \text{sprawl} & \text{loose} \\
\hline
\text{none} & 71.8 & 74.3 & 74.4 & 74.5 & 73.3 \\
\hline
\text{capital:thread} & 72.3 & 74.6 & 74.7 & 74.9 & 73.6 \\
\text{tear} & 72.4 & 74.7 & 74.7 & 74.9 & 73.6 \\
\text{sprawl} & 72.4 & 74.7 & 74.7 & 74.9 & 73.4 \\
\text{loose} & 72.4 & 74.8 & 74.7 & \textbf{74.9} & 73.3 \\
\text{strict} & \textbf{71.4} & 73.7 & 73.7 & 73.9 & 72.7 \\
\text{strict’} & \textbf{71.0} & 73.1 & 73.1 & 73.2 & 72.1 \\
\hline
\end{array}\]
to 45 words (excluding Section 23). Table 3 shows evaluation results on held-out data (all sentences), using “add-one” smoothing. All constraints other than strict improve accuracy by about a half-a-point, from 71.8 to 72.4%, suggesting that capitalization is informative of certain regularities not captured by DBM grammars; moreover, it still continues to be useful when punctuation-based constraints are also enforced, boosting accuracy from 74.5 to 74.9%.

### 5 Multi-Lingual Grammar Induction

So far, we showed only that capitalization information can be helpful in parsing a very specific genre of English. Next, we tested its ability to generally aid dependency grammar induction, focusing on situations when other bracketing cues are unavailable.

We experimented with 14 languages from 2006/7 CoNLL shared tasks (Buchholz and Marsi, 2006; Nivre et al., 2007), excluding Arabic, Chinese and Japanese (which lack case), as well as Basque and Spanish (which are pre-processed in a way that loses relevant capitalization information). For all remaining languages we trained only on simple sentences — those lacking sentence-internal punctuation — from the relevant training sets (for blind evaluation).

Restricting our attention to a subset of the available training data serves a dual purpose. First, it allows us to estimate capitalization’s impact where no other (known or obvious) cues could also be used. Otherwise, unconstrained baselines would not yield the strongest possible alternative, and hence not the most interesting comparison. Second, to the extent that presence of punctuation may correlate with sentence complexity (Frank, 2000), there are benefits to “starting small” (Elman, 1993): e.g., relegating full data to later stages helps training (Spitkovsky et al., 2010a; Cohn et al., 2011; Tu and Honavar, 2011).

Our base systems induced DBM-1, starting from uniformly-at-random chosen parse trees (Cohen and Smith, 2010) of each sentence, followed by inside-outside re-estimation (Baker, 1979) with “add-one” smoothing. Capitalization-constrained systems differed from controls in exactly one way: each learner got a slight nudge towards more promising structures by choosing initial seed trees satisfying an appropriate constraint (but otherwise still uniformly).

Table 4 contains the stats for all 14 training sets, ordered by number of multi-token fragments in training data.

---

### Table 4: Parsing performance for grammar inducers trained with capitalization-based initial constraints, tested against 14 held-out sets from 2006/7 CoNLL shared tasks, and ordered by number of multi-token fragments in training data.

| CoNLL Year & Language | Filtered Training Tokens | Sentences | Directed Accuracies with Initial Constraints | Fragments |
|-----------------------|--------------------------|-----------|---------------------------------------------|-----------|
|                        |                          |           | none thread tear sprawl loose strict strict | Multi Single |
| German 2006            | 139,333                  | 12,296    | 36.3 36.3 36.3 39.1 36.2 36.3 36.3 39.1 | 3,287 30,435 |
| English '7             | 187,505                  | 20,378    | 51.3 51.3 51.3 52.5 52.5 51.4 51.4 52.5 | 1,831 6,722 |
| Bulgarian '6           | 46,599                   | 5,241     | 59.4 59.3 59.3 59.1 59.3 59.1 59.1 59.3 | 184 1,506 |
| Czech '6               | 74,023                   | 1,599     | 21.3 17.7 22.7 21.5 21.4 31.4 27.9 31.4 | 113 317 |
| Dutch '6               | 11,943                   | 842       | 28.1 46.1 46.3 46.3 46.4 31.1 31.0 31.1 | 113 456 |
| English '7             | 72,043                   | 7,107     | 45.9 45.8 45.9 45.8 45.8 45.7 45.7 45.7 | 89 4,335 |
| Greek '7               | 8,540                    | 1,258     | 43.2 43.1 43.2 43.2 43.2 43.2 43.2 43.2 | 27 950 |
| Italian '7             | 62,811                   | 4,082     | 61.3 61.3 61.3 61.3 61.3 61.3 61.3 61.3 | 28 2,828 |
| Turkish '6             | 76,523                   | 2,835     | 32.9 32.9 32.2 33.0 33.6 33.6 33.6 33.6 | 27 950 |
| Hungarian '7           | 24,494                   | 2,042     | 68.9 67.1 69.1 69.2 68.9 68.9 68.9 68.9 | 9 950 |
| Swedish '6             | 41,918                   | 4,105     | 48.6 48.6 48.6 48.5 48.5 48.5 48.5 48.5 | 7 277 |
| Median:               | 42.5                     | 46.0      | 46.1 46.0 45.0 44.7 32.5 | 3 296 |
| Mean:                 | 42.8                     | 44.4      | 44.8 45.0 | 44.3 36.9 | 1 63 |

---

2 We used “early-stopping lateen EM” (Spitkovsky et al., 2011a, §2.3) instead of thresholding or waiting for convergence.

3 Starting from five parse trees for each sentence (using constraints thread through strict) was no better, at 44.8% accuracy.
6 Capitalizing on Punctuation in Inference

Until now we avoided using punctuation in grammar induction, except to filter data. Yet our pilot experiments indicated that both kinds of information are helpful in the decoding stage of a supervised system.

We took trained models obtained using the sprawl nudge (from §5) and proceeded to again apply constraints in inference (as in §4). Capitalization alone increased parsing accuracy only slightly, from 45.0 to 45.1%, on average. Using punctuation constraints instead led to more improved performance: 46.5%. Combining both types of constraints again resulted in slightly higher accuracies: 46.7%. Table 5 breaks down our last average performance number by language and shows the combined approach to be competitive with state-of-the-art. We suspect that further improvements could be attained by also incorporating both constraints in training and with full data.

7 Discussion and A Few Post-Hoc Analyses

Our discussion, thus far, has been English-centric. Nevertheless, languages differ in how they use capitalization (and even the rules governing a given language tend to change over time — generally towards having fewer capitalized terms). For instance, adjectives derived from proper nouns are not capitalized in French, German, Polish, Spanish or Swedish, unlike in English (see Table 1: JI). And while English forces capitalization of the first-person pronoun in the nominative case, I (see Table 1: PRP), in Danish it is the plural second-person pronoun (also I) that is capitalized; further, formal pronouns (and their case-forms) are capitalized in German (Sie and Ihre, Ihres...), Italian, Slovenian, Russian and Bulgarian.

In contrast to pronouns, single-word proper nouns — including personal names — are capitalized in nearly all European languages. Such shortest bracketings are not particularly useful for constraining sets of possible parse trees in grammar induction, however, compared to multi-word expressions; from this perspective, German appears less helpful than most cased languages, because of noun compounding, despite prescribing capitalization of all nouns. Another problem with longer word-strings in many languages is that, e.g., in French (as in English) lower-case prepositions may be mixed in with contiguous groups of proper nouns: even in surnames, the German particle von is not capitalized, although the Dutch van is, unless preceded by a given name or initial — hence Van Gogh, yet Vincent van Gogh.

7.1 Constraint Accuracies Across Languages

Since even related languages (e.g., Flemish, Dutch, German and English) can have quite different conventions regarding capitalization, one would not expect the same simple strategy to be uniformly useful — or useful in the same way — across disparate languages. To get a better sense of how universal our constraints may be, we tabulated their accuracies for the full training sets of the CoNLL data, after all grammar induction experiments had been executed.

Table 6 shows that the less-strict capitalization-induced constraints all fall within narrow (yet high) bands of accuracies of just a few percentage points: 99–100% in the case of thread, 98–100% for tear, 95–99% for sprawl and 94–99% for loose. By contrast, the ranges for punctuation-induced constraints are all at least 10%. We do not see anything partic-
7.3 Odds and Ends

Our earlier analysis excluded sentence-initial words because their capitalization is, in a way, trivial. But for completeness, we also tested constraints derived from this source, separately (see Table 2: initials). As expected, the new constraints scored worse (despite many automatically-correct single-word fragments) except for strict, whose binding constraints over singletons drove up accuracy. It turns out, most first words in WSJ are leaves — possibly due to a dearth of imperatives (or just English’s determiners).

We broadened our investigation of the “first leaf”

| CoNLL Year & Language | Total Training | Capitalization-Induced Constraints | Punctuation-Induced Constraints |
|-----------------------|----------------|-------------------------------------|---------------------------------|
|                       | Tokens / Sentences | thr-d | tear | spr-l | loose | str | strict | thr-d | tear | spr-l | loose | str | strict |
| Arabic ’06            | 52,752          | 1,460 | —   | —    | —    | —   | —     | 89.6  | 89.5 | —    | —    | —   | —     |
| ’07                   | 102,375         | 2,912 | —   | —    | —    | —   | —     | 90.9  | 90.6 | —    | —    | —   | —     |
| Basque ’06            | 41,013          | 3,190 | —   | —    | —    | —   | —     | 96.2  | 95.7 | 92.3  | 81.9 | 42.8 | 25.6  |
| ’07                   | 162,985         | 12,823 | 99.8 | 99.5 | 96.6 | 96.4 | 51.8  | 81.0  | 97.6 | 97.2 | 96.1 | 74.7 | 36.7  |
| Bulgarian ’06         | 380,525         | 14,958 | 100 | 99.5 | 95.0 | 94.6 | 15.8  | 57.9  | 96.1 | 95.5 | 94.6 | 73.7 | 36.0  |
| Catalan ’07           | 337,162         | 56,957 | —   | —    | —    | —   | —     | —    | —    | —    | —    | —   | —     |
| ’07                   | 337,175         | 56,957 | —   | —    | —    | —   | —     | —    | —    | —    | —    | —   | —     |
| Chinese ’06           | 1,063,413       | 72,703 | 99.7 | 98.3 | 96.2 | 95.4 | 42.4  | 68.0  | 89.4 | 89.2 | 87.7 | 68.9 | 37.2  |
| ’07                   | 368,624         | 25,364 | 99.7 | 98.3 | 96.1 | 95.4 | 42.6  | 67.6  | 89.5 | 89.3 | 87.8 | 69.3 | 37.4  |
| Danish ’06            | 80,743          | 5,190 | 99.9 | 99.4 | 98.3 | 97.0 | 59.0  | 69.7  | 96.9 | 96.9 | 95.2 | 68.3 | 39.6  |
| ’07                   | 172,958         | 13,349 | 99.9 | 99.1 | 98.4 | 96.6 | 16.6  | 46.3  | 89.6 | 89.5 | 86.4 | 69.6 | 42.5  |
| Dutch ’06             | 395,139         | 18,577 | 99.3 | 98.7 | 98.0 | 96.0 | 17.5  | 24.8  | 91.5 | 91.4 | 90.6 | 76.5 | 39.2  |
| English ’07           | 605,337         | 39,216 | 99.6 | 98.0 | 96.7 | 96.4 | 41.7  | 57.1  | 94.5 | 93.9 | 90.7 | 71.1 | 37.2  |
| German ’06            | 58,766          | 2,705 | 99.9 | 99.3 | 98.5 | 96.6 | 13.6  | 50.1  | 91.3 | 91.0 | 89.8 | 75.7 | 43.7  |
| Greek ’07             | 111,464         | 6,034 | 99.9 | 98.1 | 95.7 | 94.4 | 46.6  | 62.0  | 96.1 | 94.0 | 89.0 | 77.1 | 28.9  |
| Hungarian ’07         | 60,653          | 3,110 | 99.9 | 99.6 | 99.0 | 98.8 | 12.8  | 68.2  | 97.1 | 96.8 | 96.0 | 77.8 | 44.7  |
| Italian ’07           | 133,927         | 17,044 | —   | —    | —    | —   | —     | 100   | 100  | 95.4 | 89.0 | 48.9  | 63.5 |
| Japanese ’06          | 177,581         | 9,071 | 100 | 99.0 | 97.6 | 97.0 | 14.4  | 37.7  | 96.0 | 95.8 | 94.9 | 74.5 | 40.3  |
| Portuguese ’07        | 23,779          | 1,534 | 99.8 | 98.9 | 98.9 | 52.0  | 84.7  | 93.3 | 93.3 | 92.6 | 72.7 | 42.7  | 45.8 |
| Slovenian ’06         | 78,068          | 3,306 | —   | —    | —    | —   | —     | 96.5  | 96.0 | 95.2 | 75.4 | 33.4  | 40.9 |
| Spanish ’06           | 163,301         | 11,042 | 99.8 | 99.6 | 99.0 | 97.0 | 24.7  | 58.4  | 90.8 | 90.4 | 87.4 | 66.8 | 31.1  |
| Swedish ’06           | 48,373          | 4,997 | 100 | 99.8 | 96.2 | 94.0 | 22.8  | 42.8  | 99.8 | 99.7 | 95.1 | 76.9 | 37.7  |
| Turkish ’06           | 54,761          | 5,635 | 100 | 99.9 | 96.1 | 94.2 | 21.6  | 42.9  | 99.8 | 99.7 | 94.6 | 76.7 | 38.2  |
| Min                   | 100       | 99.9 | 99.0 | 98.9 | 98.9 | 59.0  | 84.7  | 100   | 100  | 96.1 | 89.0 | 48.9  | 63.5 |
| Max                   | 100       | 99.9 | 99.0 | 98.9 | 98.9 | 59.0  | 84.7  | 100   | 100  | 96.1 | 89.0 | 48.9  | 63.5 |

Table 6: Accuracies for capitalization- and punctuation-induced constraints on all (full) 2006/7 CoNLL training sets.

ularly special about Greek or Italian in these summaries that could explain their substantial improvements (18 and 11%, respectively — see Table 4), though Italian does appear to mesh best with the sprawl constraint (not by much, closely followed by Swedish). And English — the language from which we drew our inspiration — barely improved with Swedish. And English — the language from which we drew our inspiration — barely improved with capitalization-induced constraints (see Table 4) and caused the lowest accuracies of thread and strict.

These outcomes are not entirely surprising: some best- and worst-performing results are due to noise, since learning via non-convex optimization can be chaotic: e.g., in the case of Greek, applying 113 constraints to initial parse trees could have a significant impact on the first grammar estimated in training — and consequently also on a learner’s final, converged model instance. We expect the averages (i.e., means and medians) — computed over many data sets — to be more stable and meaningful than the outliers.

7.2 Immediate Impact from Capitalization

Next, we considered two settings that are less affected by training noise: grammar inducers immediately after an initial step of constrained Viterbi EM and supervised DBM parsers (trained on sentences with up to 45 words), for various languages in the CoNLL sets. Table 7 shows effects of capitalization to be exceedingly mild, both if applied alone and in tandem with punctuation. Exploring better ways of incorporating this informative resource — perhaps as soft features, rather than as hard constraints — and in combination with punctuation- and markup-induced bracketings could be a fruitful direction.
| CoNLL Year & Language | Evaluation Tokens / Sents | Bracketings capital. punct. | Unsupervised Training init. | 1-step constrained none | Supervised Parsing capital. punct. both |
|-----------------------|---------------------------|-----------------------------|-----------------------------|------------------------|--------------------------------------|
| Arabic '06            | 5,215 / 146               | —                           | 101                         | 18.4 / 20.6            | 59.8                                 |
| '07                   | 4,537 / 130               | —                           | 311                         | 19.0 / 23.5            | 63.5                                 |
| Basque '07            | 4,511 / 334               | —                           | 547                         | 17.4 / 22.4            | 58.4                                 |
| Bulgarian '06         | 5,032 / 398               | 44                         | 552                         | 19.4 / 28.9            | 76.7 / 76.8                          |
| '07                   | 4,478 / 167               | 24                         | 398                         | 18.0 / 25.1            | 78.1 / 78.2                          |
| Chinese '06           | 5,012 / 867               | —                           | 23.5                        | 27.2                   | 83.7                                 |
| '07                   | 5,161 / 690               | —                           | 19.4                        | 25.0                   | 81.0                                 |
| Czech '06             | 5,000 / 365               | 48                         | 549                         | 18.6 / 19.7            | 64.9 / 64.8                          |
| '07                   | 4,029 / 286               | 85                         | 590                         | 27.4 / 26.0            | 71.9 / 72.0                          |
| Danish '06            | 4,978 / 322               | 85                         | 590                         | 19.5 / 17.9            | 60.9 / 60.9                          |
| '07                   | 4,989 / 386               | 28                         | 318                         | 18.7 / 17.9            | 60.9 / 60.9                          |
| Dutch '06             | 4,386 / 214               | 151                        | 423                         | 17.6 / 24.0            | 65.2 / 65.6                          |
| '07                   | 4,886 / 357               | 135                        | 523                         | 23.0 / 23.7            | 70.7 / 70.7                          |
| German '06            | 4,307 / 197               | 47                         | 372                         | 17.1 / 17.1            | 71.3 / 71.6                          |
| '07                   | 6,090 / 390               | 28                         | 893                         | 17.1 / 18.5            | 67.3 / 67.2                          |
| Greek '06             | 4,360 / 249               | 71                         | 505                         | 18.6 / 32.5            | 66.0 / 65.9                          |
| '07                   | 5,005 / 709               | —                           | 26.5                        | 36.8                   | 85.1                                 |
| Hungarian '06         | 5,009 / 288               | 29                         | 559                         | 19.3 / 24.2            | 80.5 / 80.5                          |
| '07                   | 5,004 / 402               | 72                         | 785                         | 18.3 / 22.5            | 70.7 / 70.7                          |
| Italian '06           | 4,873 / 389               | 14                         | 417                         | 20.2 / 31.4            | 74.9 / 74.9                          |
| '07                   | 6,288 / 623               | 18                         | 683                         | 20.4 / 26.4            | 66.1 / 66.0                          |
| Japanese '06          | 4,991 / 206               | —                           | 453                         | 18.0 / 19.3            | 69.5                                 |
| '07                   | 5,004 / 402               | —                           | 453                         | 18.0 / 19.3            | 69.5                                 |
| Portuguese '06        | 6,090 / 390               | 28                         | 893                         | 17.1 / 18.5            | 67.3 / 67.2                          |
| '07                   | 4,886 / 357               | 135                        | 523                         | 64.4 / 24.2            | 70.7 / 70.7                          |
| Slovenian '06         | 4,307 / 197               | 47                         | 372                         | 17.1 / 17.1            | 71.3 / 71.6                          |
| '07                   | 6,090 / 390               | 28                         | 893                         | 17.1 / 18.5            | 67.3 / 67.2                          |
| Spanish '06           | 4,873 / 389               | 14                         | 417                         | 20.2 / 31.4            | 74.9 / 74.9                          |
| '07                   | 6,288 / 623               | 18                         | 683                         | 20.4 / 26.4            | 66.1 / 66.0                          |
| Swedish '06           | 4,991 / 206               | —                           | 453                         | 18.0 / 19.3            | 69.5                                 |
| '07                   | 6,288 / 623               | 18                         | 683                         | 20.4 / 26.4            | 66.1 / 66.0                          |
| Turkish '06           | 3,983 / 300               | 4                          | 305                         | 20.3 / 24.8            | 67.3                                 |
| '07                   |                               |                             |                             |                        |                                     |
| Max:                  |                            |                             |                             |                        | 80.5 / 80.5                          |
| Mean:                 |                            |                             |                             |                        | 70.9 / 70.9                          |
| Min:                  |                            |                             |                             |                        | 69.5                                 |

Table 7: Unsupervised accuracies for uniform-at-random projective parse trees (init), also after a step of Viterbi EM, and supervised performance with induced constraints, on 2006/7 CoNLL evaluation sets (sentences under 145 tokens).

phenomenon and found that in 16 of the 19 CoNLL languages first words are more likely to be leaves than other words without dependents on the left; last words, by contrast, are more likely to take dependents than expected. These propensities may be related to the functional tendency of languages to place old information before new (Ward and Birner, 2001) and could also help bias grammar induction. Lastly, capitalization points to yet another class of words: those with identical upper- and lower-case forms. Their constraints too tend to be accurate (see Table 2: uncased), but the underlying text is not particularly interesting. In WSJ, caseless multi-token fragments are almost exclusively percentages (e.g., the two tokens of 10%), fractions (e.g., 1 1/4) or both. Such boundaries could be useful in dealing with financial data, as well as for breaking up text in languages without capitalization (e.g., Arabic, Chinese and Japanese). More generally, transitions between different fonts and scripts should be informative too.

8 Conclusion
Orthography provides valuable syntactic cues. We showed that bounding boxes signaled by capitalization changes can help guide grammar induction and boost unsupervised parsing performance. As with punctuation-delimited segments and tags from web markup, it is profitable to assume only that a single word derives the rest, in such text fragments, without further restricting relations to external words — possibly a useful feature for supervised parsing models.

Our results should be regarded with some caution, however, since improvements due to capitalization in grammar induction experiments came mainly from two languages, Greek and Italian. Further research is clearly needed to understand the ways that capitalization can continue to improve parsing.
Acknowledgments

Funded, in part, by Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract FA8750-09-C-0181. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, AFRL, or the US government. We also thank Ryan McDonald and the anonymous reviewers for helpful comments on draft versions of this paper.

References

J. K. Baker. 1979. Trainable grammars for speech recognition. In *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*.

E. J. Briscoe. 1994. Parsing (with) punctuation, etc. Technical report, Xerox European Research Laboratory.

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *CoNLL*.

S. B. Cohen and N. A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *NAACL-HLT*.

S. B. Cohen and N. A. Smith. 2010. Viterbi training for PCFGs: Hardness results and competitiveness of uniform initialization. In *ACL*.

S. B. Cohen, D. Das, and N. A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *EMNLP*.

T. Cohn, P. Blunsom, and S. Goldwater. 2011. Inducing tree-substitution grammars. *Journal of Machine Learning Research*.

M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

J. L. Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48.

R. Frank. 2000. From regular to context-free to mildly context-sensitive tree rewriting systems: The path of child language acquisition. In A. Abeillé and O. Rambow, editors, *Tree Adjoining Grammars: Formalisms, Linguistic Analysis and Processing*, CSLI Publications.

K. Gimpel and N. A. Smith. 2011. Concavity and initialization for unsupervised dependency grammar induction. Technical report, CMU.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19.

D. Mareček and Z. Zabokrtský. 2011. Gibbs sampling with treeness constraint in unsupervised dependency parsing. In *ROBUS*.

R. McDonald, S. Petrov, and K. Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*.

T. Naseem and R. Barzilay. 2011. Using semantic cues to learn syntax. In *AAAI*.

T. Naseem, H. Chen, R. Barzilay, and M. Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *EMNLP*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *EMNLP-CoNLL*.

F. Pereira and Y. Schabes. 1992. Inside-outside reestimation from partially bracketed corpora. In *ACL*.

E. Ponte, J. Baldridge, and K. Erk. 2010. Simple unsupervised identification of low-level constituents. In *ICSC*.

E. Ponte, J. Baldridge, and K. Erk. 2011. Simple unsupervised grammar induction from raw text with cascaded finite state models. In *ACL-HLT*.

M. S. Rasooli and H. Faili. 2012. Fast unsupervised dependency parsing with arc-standard transitions. In *ROBUS-UNSUP*.

Y. Seginer. 2007. Fast unsupervised incremental parsing. In *ACL*.

A. Søgaard. 2011a. Data point selection for cross-language adaptation of dependency parsers. In *ACL-HLT*.

A. Søgaard. 2011b. From ranked words to dependency trees: two-stage unsupervised non-projective dependency parsing. In *TextGraphs*.

V. I. Spitkovsky, H. Alshawi, and D. Jurafsky. 2010a. From Baby Steps to Leapfrog: How “Less is More” in unsupervised dependency parsing. In *NAACL-HLT*.

V. I. Spitkovsky, D. Jurafsky, and H. Alshawi. 2010b. Profiting from mark-up: Hyper-text annotations for guided parsing. In *ACL*.

V. I. Spitkovsky, H. Alshawi, and D. Jurafsky. 2011a. Lateen EM: Unsupervised training with multiple objectives, applied to dependency grammar induction. In *EMNLP*.

V. I. Spitkovsky, H. Alshawi, and D. Jurafsky. 2011b. Punctuation: Making a point in unsupervised dependency parsing. In *CoNLL*.

V. I. Spitkovsky, A. X. Chang, H. Alshawi, and D. Jurafsky. 2011c. Unsupervised dependency parsing without gold part-of-speech tags. In *EMNLP*.

V. I. Spitkovsky, H. Alshawi, and D. Jurafsky. 2012. Three dependency-and-boundary models for grammar induction. In *EMNLP-CoNLL*.

K. Tu and V. Honavar. 2011. On the utility of curricula in unsupervised learning of probabilistic grammars. In *IJCAI*.

G. Ward and B. J. Birner. 2001. Discourse and information structure. In D. Schiffrin, D. Tannen, and H. Hamilton, editors, *Handbook of Discourse Analysis*. Oxford: Basil Blackwell.