左隅型解析に基づく文法の普遍性を利用した多言語教師なし構文解析

能地 宏　　宮尾 祐介
総合研究大学院大学 情報学専攻 / 国立情報学研究所
{noji,yusuke}@nii.ac.jp

1 はじめに

言語の文法的性質や語順は言語毎に非常に異なっているが、どれもコミュニケーションの道具であるという点は共通である。本稿の目標は、一見異なって見える言語の文法に潜む共通の性質 (文法の普遍性) を抽出し、それにより様々な言語の文法の自動獲得の性能を向上させることである。

心理言語学における有名な観察として、人間は中央埋め込みと呼ばれる構造の文を正しく処理できないことが知られている (Gibson, 2000)。例えば *The reporter who the senator who Mary met attacked ignored the president.* や「書記が代議士が首相がうたた寝したと抗議したと報告した。」は中央埋め込み文の例である。また Noji and Miyao (2014) で示したように、中央埋め込みは定量的にもツリーバンク上で多言語に渡って稀、つまり構造の難しさが言語普遍的である。

中央埋め込みと深い関連を持つ構文解析アルゴリズムとして左隅型構文解析法 (Resnik, 1992) が知られている。これはスタックの上で文の解析を行うアルゴリズムで、中央埋め込みのない全ての文をスタック深さ 1 で、一段の埋め込みを含む文はスタック深さ 2 で... と、必要なスタック深さが中央埋め込み構造にのみ増え、埋め込み度合いに応じて線形に増加する性質をもつ。

本稿はこの左隅型構文解析の文法獲得への応用可能性を探るものである。文の構造が中央埋め込みを避けやすいという傾向は言語普遍的であり、一種の文法の普遍性と考えられる。左隅型解析は木構造の中央埋め込み度合いを効率的に計算することが可能であるため、例えば文法の学習中に中央埋め込みに対してペナルティを与えることができる。本稿ではこのようなペナルティが果たして有効であるかどうかを検証する。

文法獲得といって特にここで着目するタスクは、品詞列からの教師なし依存構造の解析である。これは主に計算面での取扱いやすさによる。本タスクは有名な生成モデルが存在し (Klein and Manning, 2004)、またその学習は典型的な PCFG の EM アルゴリズムでの学習に帰着することが知られる。この PCFG において各非終端記号 X_a は主辞となる終端記号 a を指し、例えば ($\mathrm{X_V}$ ($\mathrm{X_N}$ DT N) ($\mathrm{X_V}$ V N)) は依存構造木 DT←N←V←N に対応している。90 年代に研究された句構造の PCFG による推定では各非終端記号が任意の記号であったため学習が困難であったが (de Marcken, 1999) 依存構造を用いることにより文法の書き換えルールの持つ意味を明確になり、学習の難しさが軽減される。

本稿で扱うモデルは基本的に能地ら (2015) で定式化した構造上の制約を組み込んだ DMV (the dependency model with valence) (Klein and Manning, 2004) である。以前の実験は英語の非常に簡単なものに限っていたが、本稿では多言語に渡る実験で有効性を検証する。以下、2 節で定式化についてのアイデアを述べた後、本稿で新しく導入する句の長さに対する制約について説明する。教師なし構文解析では適切な評価の仕方が問題となるが、3 節では我々の導入する構造上の制約の影響を正しく評価するための実験設定について述べる。その後 4 節で実験結果を示し、最後にまとめを行う。

2 制約つき EM アルゴリズム

本稿では次の確率モデルを考え、このモデルのパラメータ θ を EM アルゴリズムにより最適化する。

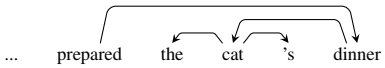
$$p'(z,x|\theta) \propto p_{\mathrm{DMV}}(z,x|\theta)f(z,x). \quad (1)$$

ここで z は構文木、x は入力の品詞列、$p_{\mathrm{DMV}}(z,x|\theta)$ は z,x に対する既存の生成モデルである DMV (後述) を表す。$f(z,x) \in [0,1]$ は構造を考慮した外部から与えられるペナルティ項であり、これが学習中のモデルにバイアスをかける。例えば能地ら (2015) で定式化した手法は、学習中に特定の深さまでの中央埋め込みを含む木構造を EM アルゴリズムの探索中に取り除くというものであるが、これは $f(z,x)$ に、z がある深さまでの埋め込みを含む場合に 0 を返すような関数を設定した場合に対応する。このような操作を可能にするのが左隅型構文解析であり、DMV などの確率モデルをこの解析アルゴリズムの上で表現し直した上でチャート型のアルゴリズムを実行すれば、式 (1) の $p_{\mathrm{DMV}}(z,x|\theta)f(z,x)$ のもとでの統計量、例えば周辺確率やルールの期待値を計算することができる。$p_{\mathrm{DMV}}(z,x|\theta)f(z,x)$ は正規化されていないが、この分布の上で EM アルゴリズムを行うことで、結果として正規化された $p'(z,x|\theta)$ の尤度を単一に上昇させられ

ることが知られている (Smith, 2006)。以上が本研究の学習の枠組みである。

DMV は依存構造木に対する単純な生成モデルであり、二種類の分布からなる。1 つは $\theta_{\mathrm{A}}(a|h, dir)$ で、これは単語 $(h,\ a)$ 間の dir 方向の依存関係の強さを決める。もう一つは単語 h がどれほど子 (dependent) を持ちやすいかを決める二値分布 $\theta_{\mathrm{S}}(stop|h, dir, adj)$ で、$stop \in \{\mathrm{STOP}, \neg\mathrm{STOP}\}$ である。$adj \in \{\mathrm{TRUE}, \mathrm{FALSE}\}$ が重要で、これは h がまだ dir 方向に子を持たない場合に TRUE となる。元々英語を意識して設計されたモデルであるが、これにより例えば、動詞は左に子 (主語) を一つだけ持ちやすい、などといった構造が捉えやすくなる。

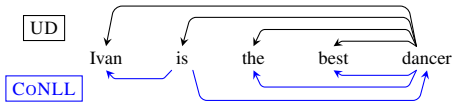
埋め込まれる句の長さによる制約　能地ら (2015) では最大スタック深さ 1 (埋め込みを許さない)、2 (一段の埋め込みのみ許す) などの制約を検討したが、経験的にこれらの中間、つまり長さの短い埋め込みのみを許す制約がうまく働くことがわかった。例えば次の構造は一段の埋め込み (the cat 's) を含むが



これはどちらかというとコーパスの単語区切りにより生じたものであると考えられ、似た文 prepared *his* dinner に対しては埋め込みは生じない。以降、最大深さ 1-L といったとき、これは長さ L までの句の埋め込みを許す。例えば上記の構造は、深さ 1-3 で許容される。これにより些細な埋め込みとそうでないものを (近似的に) 区別し、より本質的に意味のある制約を実現できるのではないかと考える。

3　実験設定

教師なし構文解析における適切な評価方法は未解決な問題の一つであり (Bisk and Hockenmaier, 2013)、これは本質的には依存構造 (主辞) の定義の仕方が一つに定まらないことに起因する。例えば以下は英語のコピュラに対する二つの主要な解析例を示したもので、



このうち CoNLL 形式 (下部) は従来の機能語に基づく解析で is を文の主辞とみなすのに対し、最近の Stanford parser などが出力する UD (Universal Dependencies) 形式 (Marneffe et al., 2014) では、依存関係は基本的に内容語の間にひかれ、文の主辞は dancer となる。この種の任意性は並列構造の解析を始めとして至るところで出現する (Zeman et al., 2014)。

教師なし構文解析の評価で最も標準的に用いられる方法は教師あり解析の場合と同様に正解の依存関係の割合に基づく UAS (unlabelled attachment score) であるが、これは現在採用している “唯一の” 正解データと比較に基づく評価であるため、上で述べた依存構造の任意性が問題となる。例えば新しい手法の導入により 10% の UAS の改善が行えたとして、これが果たして本当に意味のある改善なのか、それとも表層的なレベルで正解データが採用しているアノテーション基準に合致した木を選択しやすくなっただけなのか、ということの判断がつきにくいのである。

本研究の目標は、構造的な制約が教師なし学習に寄与する影響を議論することであるから、上記のようなアノテーション基準に起因する問題をなるべく排除した上で実験を行いたい。具体的には以下のように注意深く実験を設計し、この問題に対処する。

- 主なデータセットとして Universal Dependencies (UD) ver. 1.1 を用いる。これは多言語ツリーバンク (18 言語) の集合であり、全てのツリーバンクが統一的に UD 形式でアノテートされている。
- 各モデルの学習中の探索範囲を UD 形式に合致する構造のみに限定する。UD の基本方針は全ての機能語を内容語の子とすることであるため、全ての機能語[1]に対し子を取ることを禁止すれば、これを達成できる[2]。
- 最終的な評価は UAS に基づき行う。

UD は非常に新しいデータセットであり、これを教師なし構文解析の評価に用いる研究は本研究が初である。そのため既存手法との比較が問題となるが、この目的のために、本稿では UD の前身である Google universal treebanks (McDonald et al., 2013) を追加で用いる。その際、機能語の制約に若干の修正を加える[3]。また多くの先行研究に従い、モデルへの入力はアノテートされた品詞列のみを用い、単語の表層情報は用いない。

モデル　木構造に対する確率モデルとして、対数線形モデルによる DMV の拡張 (Berg-Kirkpatrick et al., 2010) を用いる。これはシンプルな拡張でありながら精度をよく向上させることが知られている。この上に E-step 中で制約を与え、その影響を評価する。比較する設定は以下の 3 種類にまとめられる。

- FUNC: 機能語の制約のみ。
- LEN: FUNC に加え、短い依存関係を好むようなバイアスをかける。具体的には各 h, d 番目の単語間の依存関係に対し $e^{-\gamma\cdot(|h-d|-1)}$ $(\gamma > 0)$ の負荷をかける。
- DEP: FUNC に加え、スタック深さを制限することで特定深さまでの中央埋め込みを禁止する。

[1] UD は言語共通の品詞セットを用いてタグ付けを行っている。以下の品詞を機能語とみなす。ADP, AUX, CONJ, DET, PART, SCONJ。

[2] これらは式 (1) 中の $f(z, x)$ の一部とみなすことができる。この制約は学習中にのみもうけ、テスト時には外す。実際、モデルは制約を満たす木のみを出力するように学習が行われており、テスト時の制約の有無はほとんど性能に影響を及ぼさない。

[3] 本データセットでは、機能語は ADP, CONJ, DET, PART の四種である。基本的に機能語は子を持たないが ADP が例外であり、これは従来のように句の主辞として働く。そのため ADP には逆に、最低限一つの子を持つような制約を課す。

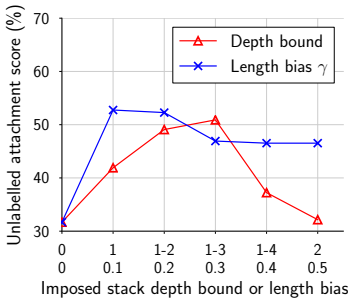


図 1: UD 版 WSJ でのパラメータ毎の精度の比較。X 軸は最大深さ (上段) または γ (下段)。

LEN は先行研究 (Smith and Eisner, 2006) で提案されたバイアスである。中央埋め込みを生じさせる依存構造は一般に長い依存関係を伴うため両者はしばしば相関するが、LEN のほうがよりシンプルであるため、これをベースラインに加えることで果たして中央埋め込みに着目する意義があるのかが議論できると考える。

文の主辞に対する制約　別の実験設定として、ベースラインに更に文法に対する (普遍的な) 知識を加えたものを用意し、その上での構造上の制約の影響を調べる。具体的には二種類の方法で文の主辞になりうる品詞に制約を設ける。Verb-or-noun 制約はこれを文中の動詞または名詞に制限する。Verb-otherwise-noun 制約は、文中に動詞がある場合動詞のみに制限し、そうでなければ名詞とする。近年の高精度な教師なし解析モデルの多くはこれらのシードとなる知識を利用するものが多い (Bisk and Hockenmaier, 2013; Naseem et al., 2010)。本稿ではこれらを試すことで、構造に対する制約と品詞についての制約がどう協業するかを議論する。

パラメータの決め方　DEP も LEN もそれぞれ一つのハイパーパラメータ (最大深さ及び γ) を持つが、本稿ではこれを英語の WSJ データを開発セットに用いて選択し、多言語データに適用する。UD の英語データは WSJ でなく Web Treebank であるが、UD 形式の WSJ は Stanford CoreNLP を用いて句構造から得ることができる。最適なパラメータは言語毎に異なる可能性が考えられるが、一つに定めることで議論がしやすくなり、またこれは多くの既存研究で使われている方針でもある (Naseem et al. (2010) など)。

4　実験

UD　まず図 1 に WSJ での結果を示す。本データセットに対しては、訓練文、テスト文ともに長さ 15 までの文に制限する。LEN について最適な値は 0.1、つまり若干のバイアスをかけたものが一番良いことがわかる。DEP については最適な値は 1-3、つまり、埋め込まれる句の長さを 3 までに制限したものが一番良い。また最大深さを 2 に設定するとベースラインとほとんど精度が変わらないことから、今回のような比較的短い文からの学習では深さ 2 の制約は緩く、あまり有効に働かないことがわかる。

	No root constraints			Verb-or-noun			Verb-otherwise-noun		
	FUNC	DEP	LEN	FUNC	DEP	LEN	FUNC	DEP	LEN
Basque	44.0	**47.6**	46.2	44.7	**54.3**	46.4	**55.8**	54.8	51.0
Bulgarian	73.6	**74.2**	70.0	73.4	**75.1**	64.1	72.7	**75.2**	70.6
Croatian	40.3	37.7	**54.5**	40.1	41.4	**47.3**	**57.0**	52.5	55.8
Czech	61.8	**64.5**	58.3	50.7	**64.7**	59.2	63.2	**66.3**	58.1
Danish	40.9	**41.3**	40.9	40.9	**41.3**	40.9	48.7	**50.1**	47.3
English	38.6	41.3	**56.6**	39.8	**41.3**	40.2	57.2	**58.5**	53.9
Finnish	26.6	25.6	**28.3**	26.2	27.7	**28.3**	40.3	34.3	**40.4**
French	35.2	**46.5**	35.8	35.7	**49.5**	47.0	44.2	**54.6**	42.1
German	49.9	**53.7**	50.4	49.7	**56.0**	51.2	49.5	**57.4**	49.9
Greek	29.0	18.3	**30.4**	61.7	**62.1**	60.2	60.5	**62.0**	60.2
Hebrew	60.7	57.1	**60.9**	52.9	**60.9**	57.5	54.8	54.2	**57.2**
Hungarian	66.7	**72.1**	63.6	68.8	**71.3**	63.6	69.2	**72.4**	64.8
Indonesian	36.4	**57.4**	50.0	32.0	**58.1**	43.6	50.2	58.5	**59.4**
Irish	64.1	**64.7**	63.0	63.1	**65.2**	63.0	63.4	**64.7**	63.9
Italian	61.2	70.8	**72.1**	62.7	**73.6**	72.5	69.2	69.8	**72.4**
Persian	**49.0**	44.7	40.6	46.9	**51.2**	39.7	48.0	**51.1**	41.7
Spanish	57.1	57.3	**62.5**	46.8	57.3	**63.1**	57.7	58.5	**62.3**
Swedish	43.4	55.2	**55.9**	**43.5**	43.2	43.5	**57.9**	53.3	56.9
Avg	48.8	51.7	**52.2**	48.9	**55.2**	51.7	56.6	**58.2**	56.0

表 1: 文の主辞への制約がある場合とない場合での多言語での精度比較。

表 1 に、この設定での多言語の結果をまとめた。まず文の主辞に制約を加えない場合 (No root constraints) DEP は多くの言語で FUNC からスコアの改善が見られるが、LEN も同様によく働き、平均してみると LEN のほうが若干精度が高い。興味深いことに、文の主辞に制約を加えると結果は異なり、DEP の精度向上が大きくなるのに対し LEN ではほとんど向上が見られなくなる。特に verb-or-noun 制約を加えた場合、DEP は 18 言語中 14 言語で最高精度を示す。文の主辞が名詞である文も存在するため verb-otherwise-noun 制約は再現率を下げる可能性があるが、結果をみると全体の精度は向上し、特にベースラインである FUNC の性能向上が著しい。これは大まかには文の主辞は動詞が大半を占めているためだと思われる。この場合でも平均の精度は DEP が最も高い。

分析　先ほどの結果から、どうも DEP と LEN は言語の異なる構造を抽出しているように見える。この違いを理解するため、英語で主辞の制約を加えない場合の両者の出力結果を比較してみたところ、エラーの傾向に興味深い違いが見られた。具体的には、DEP は正しい句の範囲を同定するのに役立つのに対し、LEN はより局所的な品詞間の関係、例えば動詞から名詞への依存関係などを正しく認識することが多い。典型的なものを図 2 に示す。“On ... pictures” に着目すると、DEP は名詞句 “the ... pictures” および前置詞句 “On ... pictures” の範囲の同定を正しく行っているのに対し、LEN の解析は句 “On ... two” が pictures に係る、というものであり、これは言語の構造を正しく捉えていない。逆に LEN は he ← look など基本的な依存関係の獲得に成功しているのに対し、DEP では逆向きの関係

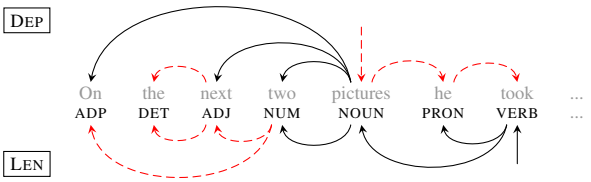


図 2: 英語 (主辞制約なし) での DEP と LEN の典型的なエラーのパターンの比較。実線は正解を表す。

が学習されてしまっている。このような違いを生んだ原因に対する一つの仮説は、DEP は中央埋め込みに対する制約、すなわち本質的には句構造に対する制約であるため、学習された依存構造のモデルも正しい句を認識しやすくなるのではないか、というものである。逆に LEN は句とは関係なしに短い依存関係を好むが、図 2 のように文の主辞が動詞であることなどを獲得するためには、より単純なこちらのほうが有効である可能性がある。

また多言語を通してみると、DEP による制約は文の主辞に対する制約と直行しやすく、そのために verb-or-noun での大幅な性能の向上が見られたのではないかと考えられる。それに対し LEN が捉える局所的な構造はこの制約と重複しやすいのではないだろうか。

Google treebanks 最後に Google treebanks 上での既存研究との比較を表 2 に示す。N10 (Naseem et al., 2010) と GE15 (Grave and Elhadad, 2015) は現在の最高精度の手法で、どちらも事後分布正則化により多数 (12 個) の品詞間のルールをモデルに組み込んでいる。先行研究に従い、本データセットに対しては訓練文の最大長さを 10 に設定した。また WSJ でのパラメータ選択を行い最大深さ 1-2, $\gamma = 0.1$ を得た。本稿の FUNC はよりも少ないルール (機能語の 4 個と文の主辞の 2 個の計 6 個) を制約として用いるが、N10 よりも高い性能を示す。そこから DEP による性能向上が大きく、全体として LEN を上回る。GE15 は最新の手法で、識別クラスタリングに基づく手法である。DEP は平均してこれよりも若干低い精度であるが、Korean 以外ではほぼ同等の性能を示している。

5 おわりに

依存構造をもとに、左隅型解析を利用した文法獲得の可能性について検討を行った。特に短い句の埋め込みのみを許す制約が有効であることを発見し、これが単純な長さに基づく制約では捉えられない構造を捉え、他の品詞間のルールと組み合わせれば機械学習的により高度な最新の手法に匹敵する性能を持ちうることを示した。この設定はまた、かなり記憶容量の制限された学習者の近似とみなすことができ、認知的観点からも興味深い結果である。最後に、本稿では依存構造の学習に焦点を当てたが、中央埋め込み及び左隅型構文解析という考えはこれに限定されるものではない。例

	FUNC	DEP	LEN	N10	GE15
	≤ 10 / $\leq \infty$	≤ 10 / $\leq \infty$	≤ 10 / $\leq \infty$	≤ 10	≤ 10 / $\leq \infty$
German	**64.5** / 49.5	64.3 / 50.6	62.5 / 49.0	53.4	60.2 / –
English	57.9 / 44.4	59.5 / 45.6	56.9 / 43.3	66.2	**62.3** / –
Spanish	68.2 / 55.5	71.1 / 58.8	69.6 / 55.5	**71.5**	68.8 / –
French	69.2 / 55.9	69.6 / 57.9	66.4 / 55.4	54.1	**72.3** / –
Indonesian	66.8 / 54.5	67.4 / 58.1	66.7 / 54.8	50.3	**69.7** / –
Italian	43.9 / 32.9	**67.3** / 58.4	44.0 / 32.8	46.5	64.3 / –
Japanese	47.5 / 43.5	54.5 / 53.7	47.6 / 43.6	**58.2**	57.5 / –
Korean	28.6 / 25.7	30.7 / 28.4	43.2 / 41.8	48.8	**59.0** / –
Portuguese	63.0 / 53.2	67.1 / 57.9	62.6 / 52.8	46.4	**68.3** / –
Swedish	67.4 / 52.1	**67.9** / 52.4	66.4 / 51.5	64.3	66.2 / –
Avg	57.7 / 46.7	62.0 / 52.2	58.6 / 48.0	56.0	**64.8** / 55.8

表 2: Google treebanks での既存研究との精度比較 (Naseem et al. (2010) 及び Grave and Elhadad (2015))。10 と ∞ はテスト時の最大文長である。本稿の手法は verb-otherwise-noun 制約を加えたもの。

えば 1 節で述べた句構造の PCFG がうまくいかなかった原因の一つは、足がかりとなる文法に対する制約が少なすぎたことであると考えられ、本稿の手法が一つの解決策となりうる。また近年、少量の事前知識を与えることで品詞列から CCG の学習が可能であることが示されている (Bisk and Hockenmaier, 2013)。今後の可能性の一つとして、本稿で導入したような構造上のバイアスにより、より少ない人手の知識からこれらの文法が獲得できるようになるかもしれない。このような検討は今後の課題である。

参考文献

T. Berg-Kirkpatrick, A. Bouchard-Côté, J. DeNero, and D. Klein. Painless Unsupervised Learning with Features. In *NAACL*, June 2010.

Y. Bisk and J. Hockenmaier. An HDP Model for Inducing Combinatory Categorial Grammars. *TACL*, 1:75–88, 2013.

C. de Marcken. On the Unsupervised Induction of Phrase-Structure Grammars. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Technology*, pages 191–208. Springer Netherlands, 1999.

E. Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126, 2000.

E. Grave and N. Elhadad. A convex and feature-rich discriminative approach to dependency grammar induction. In *ACL*, July 2015.

D. Klein and C. Manning. Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency. In *ACL*, July 2004.

M. D. Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C. D. Manning. Universal stanford dependencies: a cross-linguistic typology. In *LREC*, May 2014.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Bertomeu Castelló, and J. Lee. Universal Dependency Annotation for Multilingual Parsing. In *ACL*, August 2013.

T. Naseem, H. Chen, R. Barzilay, and M. Johnson. Using Universal Linguistic Knowledge to Guide Grammar Induction. In *EMNLP*, October 2010.

H. Noji and Y. Miyao. Left-corner Transitions on Dependency Parsing. In *COLING*, August 2014.

能地 宏, 宮尾 祐介, M. Johnson and J. Pate. 左隅型解析を利用した無情報からの教師なし係り受け解析. 言語処理学会 第 *21* 回年次大会, pages 401–404, 2015.

P. Resnik. Left-Corner Parsing And Psychological Plausibility. In *COLING*, pages 191–197, 1992.

N. A. Smith. *Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text*. PhD thesis, Johns Hopkins University, October 2006.

N. A. Smith and J. Eisner. Annealing Structural Bias in Multilingual Weighted Grammar Induction. In *COLING-ACL*, pages 569–576, July 2006.

D. Zeman, D. Dušek, O.and Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637, 2014.