Morphological and Syntactic Case in Statistical Dependency Parsing

Wolfgang Seeker*
University of Stuttgart

Jonas Kuhn**
University of Stuttgart

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Most morphologically rich languages with free word order use case systems to mark the grammatical function of nominal elements, especially for the core argument functions of a verb. The standard pipeline approach in syntactic dependency parsing assumes a complete disambiguation of morphological (case) information prior to automatic syntactic analysis. Parsing experiments on Czech, German, and Hungarian show that this approach is susceptible to propagating morphological annotation errors when parsing languages displaying syncretism in their morphological case paradigms. We develop a different architecture where we use case as a possibly underspecified filtering device restricting the options for syntactic analysis. Carefully designed morpho-syntactic constraints can delimit the search space of a statistical dependency parser and exclude solutions that would violate the restrictions overtly marked in the morphology of the words in a given sentence. The constrained system outperforms a state-of-the-art data-driven pipeline architecture, as we show experimentally, and, in addition, the parser output comes with guarantees about local and global morpho-syntactic wellformedness, which can be useful for downstream applications.

1. Introduction

In statistical parsing, many of the first models were developed and optimized for English. This is not surprising, given that English is the predominant language for research in both computational linguistics and linguistics proper. By design, the statistical parsing approach avoids language-specific decisions built into the model architecture; models should in principle be trainable on any data following the general treebank representation scheme. At the same time, it is well known from theoretical and typological work in linguistics that there is a broad multi-dimensional spectrum of language types, and that English is in a rather "extreme" area in that it marks grammatical relations (subject, object, etc.) strictly with phrase-structural configurations. There are only residues of an inflectional morphology left. In other words, one

* Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Pfaffenwaldring 5b, D-70569 Stuttgart, Germany. E-mail: seeker@ims.uni-stuttgart.de.
** Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Pfaffenwaldring 5b, D-70569 Stuttgart, Germany. E-mail: jonas@ims.uni-stuttgart.de.

Submission received: 30 September 2011; revised submission received: 20 May 2012; accepted for publication: 3 August 2012.

© 2013 Association for Computational Linguistics

cannot exclude that architectural or representational modeling decisions established as empirically useful on English data may be favoring the specific language type of English. Indeed, carrying over successful model architectures from English to typologically different languages mostly leads to a substantial drop in parsing accuracy. Linguistically aware representational adjustments can help reduce the problem significantly, as Collins et al. (1999) showed in their pivotal study adjusting a statistical (constituent) parsing model to a highly inflectional language with free word order, Czech in that case, pushing the results more than seven percentage points up to a final 80% dependency accuracy (as compared with 91% accuracy for the English "source" parser on the *Wall Street Journal*). Even in recent years, however, a clear gap has remained between the top parsing architecture for English and morphologically rich(er) languages.[1] The relative hardness of the parsing task, compared with English, cuts across statistical parsing approaches (constituent or dependency parsing) and across morphological subtypes, such as languages with a moderately sized remaining inflectional system (like German), highly inflected languages (like Czech), and languages in which interactions with derivational morphology make the segmentation question non-trivial (such as Turkish or Arabic, compare, for example, Eryiğit, Nivre, and Oflazer [2008]).

Still, it remains hard to pinpoint systematic architectural or representational factors that *explain* the empirical picture, although there is a collection of "recipes" one can try to tune an approach to a "hard language." Of course, there are good reasons for adjusting a well-proven system rather than developing a more general one from scratch—given that part of the success of statistical parsing in general lies in subtle ways of exploiting statistical patterns that reflect inaccessible levels of information in an indirect way.

This article attempts to do justice to the special status of mature data-driven systems and still contribute to a systematic clarification, by (1) focusing on a clear-cut aspect of morphological marking relevant to syntactic parsing (namely, case marking of core arguments); (2) comparing a selection of languages covering part of the typological spectrum (Czech, German, and Hungarian); (3) using a state-of-the-art data-driven parser (Bohnet 2009, 2010) to establish how far the technique of representational adjustments may take us; and (4) performing a problem-oriented comparison with an alternative architecture, which allows us to add constraints motivated from linguistic considerations.

In a first experiment, we vary the morphological information available to the parser and examine the errors of the parser with respect to the case-related functions. It turns out that although the parser is indeed able to learn the case-function mapping for all three languages, it is susceptible to errors that are propagated through the pipeline model when parsing languages that show syncretism[2] in their morphological paradigms, in our case Czech and German (e. g., for neuter nouns, nominative and accusative case have the same surface form). In contrast, due to its mostly unambiguous case system, we find a much smaller effect for Hungarian. Although the parser itself profits much from morphological information as our experiments with gold standard morphology show, errors in automatically predicted morphological information frequently cause errors in the syntactic analysis.

1 Compare, for example, the various Shared Tasks on parsing multiple languages, such as the CoNLL Shared Tasks 2006, 2007, 2009 (Buchholz and Marsi 2006; Nivre et al. 2007a; Hajič et al. 2009), or the PaGe Shared Task on parsing German (Kübler 2008).

2 Two or more different feature values are signaled by the same form.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

In order to better handle syncretism in the morphological description, we then propose a different way of integrating morphology into the parsing process. We develop an alternative architecture that circumvents the strict separation of morphological and syntactic analysis in the pipeline model. We adopt the integer linear programming (henceforth ILP) approach by Martins, Smith, and Xing (2009), which we augment with a set of linguistically motivated constraints modeling the morpho-syntactic dependencies in the languages. Case is herein interpreted as an underspecified filtering device that guides a statistical model by restricting the search space of the parser. Due to the constraints, the output of the ILP parser is guaranteed to obey all syntactic restrictions that are marked overtly in the morphological form of the words. Although the restrictions are implemented as symbolic constraints, they are applied to the parser during the search for the best tree, which is driven by a statistical model. We show in a second experiment that restricting the search space in this way improves the performance on argument functions (indicated by case morphology) considerably on all three languages while the performance on all other functions stays stable.

We proceed by first discussing the role of case morphology in syntax (Section 2), followed by a presentation of the parsing architecture of the Bohnet parser with a discussion of the relevant aspects for our first experiment (Section 3). Next, we compare the morphological annotation quality of automatic tools with the gold standard across languages (Section 4). We then turn to the first experiment in this article where we examine the performance of the parser with respect to core argument functions on the three languages (Section 5). In the second experiment (Section 6), we apply an ILP parser to the data sets augmented with a set of linguistic constraints that integrate morphological information in an underspecified way into the parsing architecture. We conclude in Section 7.

2. Challenges of Parsing Morphologically Rich Languages

A characteristic property of most languages commonly referred to as *morphologically rich* is that they use morphological means at the word level to encode grammatical relations within the sentence rather than using the phrase-structural configuration. Whereas in English or Chinese, placement of a word (or phrase) in a particular position relative to the verbal head marks its function (e. g., as the subject or object), morphologically rich languages encode grammatical relations largely by changing the morphological form of the dependent word, the head word, or both. A correlated phenomenon is the free word order for which many of these languages allow. Because information about grammatical relations is marked on the words themselves, it stays available regardless of their relative position, so word order can be used to mark other information such as topic-focus structure. The richer the morphological system, the freer the word order tends to be, or, as Bresnan (2001) puts it, *morphology competes with syntax*. We thus see that typologically, morphological and syntactic systems are interdependent and influence each other. Most languages are located somewhere along a continuum between purely configurational and purely morphological marking.

In principle, data-driven parsing models with word form sensitive features have the potential to not only pick up configurational patterns for grammatical relation marking, but also systematic patterns in the observed variation and co-variation of morphological word forms. It is, however, not only the interaction between syntax and morphology that adds challenges—the marking patterns are also non-trivial to pick up from surface data.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

One of the linguistic challenges is that there are different, overlapping regimes for morphological marking. One can distinguish head-marking and dependent-marking of a grammatical relation, depending on where the inflection occurs. In addition, Nichols (1986, page 58) identifies four ways in which inflection markers may play a role in signaling syntactic dependency:

Example 1
Hebrew, taken from Nichols (1986, page 58)

bēt *sefer*
house-of book
'school', lit. 'book house'

First, the morphological marker simply registers the *presence* of a syntactic dependency. In Example (1), the form of the word *bēt* signals the presence of a dependent, without specifying the nature of the relation.

Second, the affix marks not only the presence but also the *type* of the dependency. A typical example of the dependent-marking kind is nominal case: Accusative case on a noun marks it not only as a dependent of a verb, but it also marks the type of relation, namely, direct object. Verb agreement markers in Indo-European languages are a head-marking kind of example: They indicate that a noun stands specifically in the subject relation. Third, a morphological marker may, in addition, *index* certain lexical or inflectional categories of the dependent on the head (or vice versa). Subject agreement often indexes the dependent subject's gender and number properties on the head verb; attributive adjectives in Czech, for instance, agree with their noun heads in case, number, and gender. Fourth, for some affixes, there is a paradigm for indexing *internal* properties of the head on the head itself (e. g., tense or mood of a verb) or properties of the dependent on the dependent (e. g., gender marking on nouns).

An additional linguistic challenge in learning the patterns from data, which we will discuss in detail in Section 2.2, comes from the fact that the inflectional paradigms may contain syncretism. This may interfere with the learning of the previously discussed patterns. Further challenges we do not address in this article include interactions between syntax and derivational morphology, which for some languages like Turkish and Arabic can go along with segmentation issues.

The first Workshop on Statistical Parsing of Morphologically Rich Languages has set the agenda for developing effective systems by identifying three main types of technical challenges (Tsarfaty et al. 2010, page 2), which we rephrase here from our system perspective:

Architectural challenges. Should data-driven syntactic parsing be split into subtasks, and how should they interact? Specifically, should morphological analysis (and likewise tokenization, part-of-speech tagging, etc.) be performed in a separate (data-driven?) module and how can error propagation through the pipeline be minimized? Can a joint model be trained on data that captures two-way interactions between several levels of representation? Should the same system modularization be used in training and decoding, or can decoding combine locally trained models, taking into account more global structural and representational constraints?

Representational challenges. At what technical level should morphological distinctions be represented? Should they (or some of them) be included at the part-of-speech (POS) level, or at a higher level in the structure? Can some type of representation help

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

avoid confusions due to syncretisms? What is the most effective set of dependency labels for capturing morphological marking of grammatical relations?

Lexical challenges. How can lexical probabilities be estimated reliably? The main problem for morphologically rich languages is the many different forms for one lexeme, which is amplified by the often limited amount of training data. How can a parser analyze unseen word forms and use the information profitably?

2.1 Previous Work and Our Approach

The first two types of technical challenges often go hand in hand, as a change in architecture effectively means a change in the interface representations, and vice versa. Collins et al. (1999) reduce the tag set for the Czech treebank, which consists of a combination of POS tags and a detailed morphological specification, in order to tackle data sparseness. A combination of POS and case features turns out to be best for their parsing models. In statistical constituent parsing, many investigations devise treebank transformations that allow the parsing models to access morphological information higher in the tree (Schiehlen 2004; Versley 2005; Versley and Rehbein 2009). These transformations apply category splits by decorating category symbols with morphological information like case. Whereas these approaches change traditional models to cope with morphological information, others approach the problem by devising new models tailored to the special requirements of morphologically rich languages. Tsarfaty and Sima'an (2008, 2010) introduce an additional layer into the parsing process that directly models the subcategorization of a non-terminal symbol without taking word order into consideration. The parser thus separates the functional subcategorization of a word from its surface realization, which is not a one-to-one relation in morphologically rich languages with free word order. In statistical dependency parsing, morphological information is mostly used as features in the statistical classifier that guides the search for the most probable tree (Bohnet 2009; Goldberg and Elhadad 2010; Marton, Habash, and Rambow 2010). The standard way established in the CoNLL Shared Tasks (Nivre et al. 2007a; Hajič et al. 2009) is a pipeline approach where POS and morphological information is predicted as a preprocessing step to the actual parsing. Although Goldberg and Elhadad (2010) and Marton, Habash, and Rambow (2010) find improvements for hand-annotated (gold) morphological features, automatically predicted morphological information has none or even negative effects on their parsing models. Goldberg and Elhadad (2010) also show that linguistically grounded, carefully designed features (here agreement between adjectives and nouns in Hebrew) can contribute a considerable improvement, however. Finally, the pipeline approach itself can be questioned. Cohen and Smith (2007), Goldberg and Tsarfaty (2008), and Lee, Naradowsky, and Smith (2011) present joint models where the processes of predicting morphological information and syntactic information are performed at the same time. All three approaches acknowledge the fact that syntax and morphology are heavily intertwined and interact with each other.

Our attempt at tackling the technical and linguistic challenges can be characterized as follows: In Section 6, we propose a system architecture that at the basic level follows a pipeline approach, where local data-driven models are used to predict the highest scoring output in each step. But this pipeline is complemented with a knowledge-based component modeling grammatical knowledge about inflectional paradigms and morphological marking of grammatical relations. Both parts are combined using a set of global constraints that model the language-specific morpho-syntactic dependencies to which a syntactic structure in that language has to adhere. These constraints are

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

used to weed out linguistically implausible structures among the candidate outputs of the parser. Our architecture thus resides between a strict pipeline approach where no step can influence previous results, and a full joint model, where several subtasks are predicted simultaneously. Using the global constraints we can precisely define the parts of the structure where an interaction between the morphology and the syntax is allowed to take place.

The key design tasks are of a representational nature: What are the linguistic units for which hard constraints can (and should) be enforced in a language? (For example, within Czech and German nominal phrases, indexing of case, number, and gender follows a strict regime—the values have to co-vary.) What underspecified interface representation is appropriate to negotiate between the potentially ambiguous output of one local component and the assumed input of another component? How can we restrict them as much as possible without sacrificing the correct solution? As it turns out, the explicit enforcement of conservative linguistic constraints over morphological and syntactic structures in decoding leads to significantly improved parsing performance on case-bearing dependents, and also to improved overall performance over a state-of-the-art data-driven pipeline approach.

2.2 Case Between Morphology and Syntax

In this article, we concentrate on the case feature, which resides at the interface between morphology and syntax. The case feature overtly marks (when unambiguous) the syntactic function of a nominal element in a language. Languages show different sophistication in their case systems. Where German has four different case values, Hungarian uses a complex system of about 20 different values. In all languages with a case system, it is used to distinguish and mark the function of the different arguments of verbs (Blake 2001). Correctly recognizing the argument structure of verbs is one of the most important tasks in automatic syntactic analysis because verbs and their arguments encode the core meaning of a sentence and are therefore essential to every subsequent semantic analysis step.

The three languages investigated in this article, German, Czech, and Hungarian, belong to the broad category of morphologically rich languages. Syntactically, they all use a case system to mark the function of the arguments of a verb (and a preposition). The morphological realization of these case systems show important differences, however, which have a direct influence on syntactic analysis. Czech and German are both Indo-European languages, Czech from the Slavonic branch and German from the Germanic branch. Hungarian, on the other hand, is a Finno-Ugric language of the Ugric branch. Czech and German both are **fusional** languages, where nominal inflection suffixes signal gender, number, and case values simultaneously. Hungarian is an **agglutinating** language, namely, every morphological feature is signaled by its own morpheme, which is appended to the word. Whereas Hungarian has a mostly unambiguous case system, Czech and (more so) German show a considerable amount of syncretism in their nominal inflection. It is this syncretism that makes it so much harder for a statistical system to learn the morphological marking patterns of a language.

Table 1 shows examples of declension paradigms for the fusional languages Czech and German. Note that these are only examples and cannot represent the entire complexity of the systems. We use them here to exemplify the widespread morphological syncretism in these two languages. In the masculine animate noun of Czech *bratr* ('brother'), ACC/GEN SG, DAT/LOC SG, and ACC/INS PL use the same word forms

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Table 1
Examples of nominal declension paradigms in Czech and German. German never distinguishes gender in plural.

Czech, masc. animate noun *brother*			**Czech**, neuter noun *city*		
MASC ANI	SG	PL	NEUT	SG	PL
NOM	bratr	bratř**i**	NOM	měst**o**	měst**a**
ACC	bratr**a**	bratr**y**	ACC	měst**o**	měst**a**
DAT	bratr**ovi/u**	bratr**ům**	DAT	měst**u**	měst**ům**
GEN	bratr**a**	bratr**ů**	GEN	měst**a**	měst
VOC	bratř**e**	–	VOC	měst**o**	–
LOC	bratr**ovi/u**	bratr**ech**	LOC	měst**ě/u**	měst**ech**
INS	bratr**em**	bratr**y**	INS	měst**em**	měst**y**

German, definite determiner *the*					**German**, masculine noun *dog*			**German**, feminine noun *woman*		
	MASC	NEUT	FEM	PL		MASC	PL		FEM	PL
NOM	der	das	die	die	NOM	Hund	Hund**e**	NOM	Frau	Frau**en**
ACC	den	das	die	die	ACC	Hund	Hund**e**	ACC	Frau	Frau**en**
DAT	dem	dem	der	den	DAT	Hund	Hund**en**	DAT	Frau	Frau**en**
GEN	des	des	der	der	GEN	Hund**es**	Hund**e**	GEN	Frau	Frau**en**

respectively.[3] In the neuter noun *město* ('city') we find syncretism in NOM/ACC SG and PL, and DAT/LOC SG. The NOM/ACC syncretism in neuter nouns is a typical property of Indo-European languages (Blake 2001). Note also that some inflection morphemes fill different paradigm cells, for instance *bratra* is ACC SG, *města* is NOM PL. To resolve the ambiguity, gender and number features need to be considered.

Unlike Czech, German has determiners, which are also marked for case and agree with their head noun in the so-called phi-features (gender, number, case). The declension patterns of determiners and nouns in German have developed in different ways, leading to highly case-ambiguous forms for nouns. We see in Table 1 two German nouns, a masculine one and a feminine one. Although the declension paradigm of the masculine noun has kept some residual formal marking of case in the GEN SG and the DAT PL, the declension pattern of the feminine noun does not show case distinction at all. Both nouns, however, mark the number feature overtly. The paradigm of the determiner is much less ambiguous in the case dimension, but shows syncretism between different number and gender features. Eisenberg (2006) calls the distribution of different kinds of syncretism over different parts of the German noun phrase *Funktionsteilung* ('function sharing'). It makes the morphological agreement between German nouns and their dependents extremely important because only by agreement can a mutual disambiguation take place and reduce the morpho-syntactic ambiguity for the noun phrase. We will show that for the fusional languages Czech and German, automatic morphological analyzers have problems predicting the correct case, number, and gender values, whereas for the agglutinating language Hungarian, the unambiguous case paradigm makes case prediction extremely easy.

3 NOM: nominative, GEN: genitive, DAT: dative, ACC: accusative, LOC: locative, INS: instrumental, SG: singular, PL: plural, M/MASC: masculine, F/FEM: feminine, N/NEUT: neuter.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

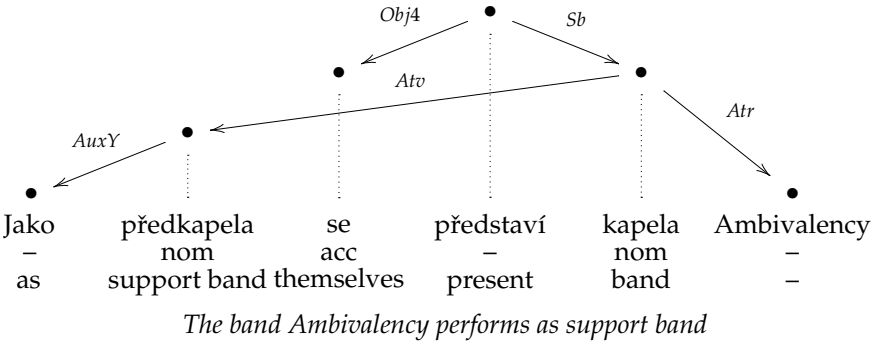

Figure 1
A dependency tree from the Czech treebank. Sentence no. 3,159 in the CoNLL 2009 data set.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

In order to see the influence of morphology on today's data-driven systems for syntactic analysis, we investigate the performance of a state-of-the-art dependency parser (Bohnet 2009, 2010) on the three languages just described paying special attention to the handling of the core grammatical functions (i. e., the argument functions of verbs). Dependency syntax (Hudson 1984; Mel'čuk 1988) models the syntactic structure of a sentence by directed labeled links between the words (tokens) of a sentence. Figure 1 shows an example tree for a Czech sentence. Every word of the tree is attached to exactly one other word (its head) by a labeled arc whose label specifies the nature of the relation. For instance, *kapela* is labeled as subject (Sb) of the sentence. Morphologically, the subject is marked with nominative case (nom) whereas the direct object (Obj4) is marked with accusative case (acc). We see that the object can precede the verb. Syntactically, Czech allows for all permutations of subject, object, and verb (Janda and Townsend 2000, page 86). It is thus a free word order language. Another property of free word order languages is the higher amount of non-projective structures (compared with English). Non-projective structures are indicated by crossing branches in the tree structure, as between *kapela* and *předkapela* in Figure 1.

3. Parsing Architecture

In this section, we give a brief description of the parser that we use in the first experiment, where we analyze the performance of the parser with respect to morphological information. The parser is the state-of-the-art data-driven second-order graph-based dependency parser presented in Bohnet (2010).[4] It is an improved version of the parser described in Bohnet (2009), which ranked first for German and second for Czech for syntactic labeled attachment score in the CoNLL 2009 Shared Task (Hajič et al. 2009).

The parser follows the standard pipeline approach. Information about lemma, POS, and morphology is automatically predicted and fully disambiguated prior to the parsing step. The CoNLL 2009 Shared Task used a tabbed format where every token in a sentence is represented by a line of tabulator-separated fields holding gold standard and predicted information about word position, word form, lemma, POS, morphology, attachment, and function label. Figure 2 gives an example for the word *se* in the sentence in Figure 1. Note that for every type of information, the

4 `http://code.google.com/p/mate-tools`, version: `anna-2`.

```
3 se se se P P SubPOS=7|Num=X|Cas=4 SubPOS=7|Num=X|Cas=4 4 4 Obj4 Obj4
```

Figure 2
Example of the CoNLL 2009 dependency format for *se*. Columns are from left to right: Position, word form, gold lemma, predicted lemma, gold POS, predicted POS, gold morphology, predicted morphology, gold head position, predicted head position, gold function label, predicted function label. Semantic information is not displayed. The gold standard columns are used for evaluation purposes.

human-annotated gold standard and the predicted value by an automatic tool is represented. The morphology columns contain several morphological features separated by a vertical bar.

The parser itself consists of two main modules, the decoder and the feature model. It is a maximum-spanning-tree[5] parser (McDonald et al. 2005; McDonald and Pereira 2006) that searches for the best-scoring tree using a chart-based dynamic programming approach similar to the one proposed by Eisner (1997). The substructures are scored by a statistical feature model that has been trained on treebank data; the best-scoring tree is the tree with the highest sum over the scores of all substructures in the tree. The actual implementation is derived from the decoder by Carreras (2007), which was shown to be efficient even for very rich feature models (Carreras 2007; Johansson and Nugues 2008; Bohnet 2009).

The features used in the statistical model are combinations of basic features, namely, word form, lemma, POS, and morphological features. In addition, the distance between two nodes, the direction of the edge, and the words between head and dependent are included. Every feature is combined with the function label on the edge. A detailed description of the feature model is beyond the scope of this article, but the interested reader can find it in Bohnet (2009, 2010).

Because we are interested in the way the parser handles morphological information, we will briefly discuss the inclusion of morphological features as described in Bohnet (2009, page 3). The parser computes morphological features by combining the part-of-speech tags (pos) of the head and the dependent with the cross-product of their morphological feature values. For this, the morphological information (see Figure 2: columns 7 and 8) is split at the vertical bar and every single morphological feature value is treated as one morphological feature in the statistical model. The cross-product then pairs the single feature values of dependent and head creating all combinations. One single feature computed for the edge between an adjective and a noun in Czech may then look like (A,N,acc,acc), which states the information that both words have the accusative case. Other features are created as well, however, that might look like (A,N,sg,masc), which states that the adjective has singular number and the noun has masculine gender. So the algorithm does not pay attention to category classes. Furthermore, the whole cross-product is computed for every edge in the tree. All features are additionally combined with the function label between the head and the dependent, so in the parsing features, a morphological feature like case is directly combined with the function label with which it appears together in the treebank. Because of this, the parser should have direct access to the information about which case value signals a particular grammatical function. Intuitively, the statistical model should learn that certain dependent head configurations often occur with certain morphological feature

5 Or graph-based as opposed to transition-based (Nivre et al. 2007b; Bohnet 2011).

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

combinations. For example, a subject edge between a noun and a verb should very often occur together with morphological features involving nominative case, and a dative object edge should often occur with a dative feature.

The statistical model is a linear multi-class classifier, trained using an on-line learning procedure (MIRA [Crammer et al. 2003] with a hash kernel [Bohnet 2010]). Learning is an iterative process where the parser repeatedly tries to recreate the training corpus sentence by sentence. If the parser makes no mistakes, it proceeds to the next training instance. Otherwise, the feature weights for the tree that would have been correct and the feature weights for the tree produced by the parser are compared and the weights in the feature model are adjusted to favor the correct tree and disfavor the incorrect one. The parser repeatedly parses the treebank, adjusting its feature model to produce trees that match the trees in the training data. Because the decoder can only derive projective trees (without crossing edges), the parser reattaches individual edges in the tree in a post-processing step to allow for non-projective trees (crossing edges, see Figure 1) using the algorithm in McDonald and Pereira (2006).

4. Data

Before we turn to our first experiment and its analysis, we briefly describe the data sets that we used in the experiments and discuss the quality of the morphological annotation. In a pipeline architecture, where morphological features are fully disambiguated prior to parsing, low quality in the predicted morphological information will have considerable impact on the ability of the parser to learn the mapping between case and grammatical functions that we want it to learn. Furthermore, the errors made in the morphological preprocessing are the first observable difference between the two fusional languages and the agglutinating language and directly reflect this typological difference. We will thus show that whereas the morphological preprocessing for Czech and German makes mistakes because of the syncretism in the morphological paradigms, the morphological preprocessing for Hungarian suffers from a different problem.

All the data sets come from the newspaper domain. The Czech data set is the CoNLL 2009 Shared Task data set consisting of 38,727 sentences from the Prague Dependency Treebank (Böhmová et al. 2000; Hajič et al. 2006). The German data set (Seeker and Kuhn 2012) is a semi-automatically corrected recreation of the data set that was used in the CoNLL 2009 Shared Task (36,017 sentences). It uses the exact[6] same raw (surface) data but contains a different syntactic annotation. It was semi-automatically derived from the original TIGER treebank (Brants et al. 2002) and some time was spent on manually correcting incorrect function labels and POS tags. The Hungarian data consist of the general newspaper subcorpus (10,188 sentences) of the Szeged Treebank (Csendes, Csirik, and Gyimóthy 2004), which was converted from the original constituent structure annotation to dependency annotation and manually checked by four trained linguists (Vincze et al. 2010). For the experiments in the following sections, we use the training splits for Czech and German, and the whole set for Hungarian.

For the Czech and the Hungarian data, we kept the predicted information for lemmata, POS, and morphology that was already provided with the data. For both

6 Except for three sentences that for some reason were missing in the 2006 version of the TiGer treebank, from which this corpus was derived. The original data set in the CoNLL 2009 Shared Task was derived from the 2005 version, which still contains these three sentences. The 2005 version also contained spelling errors in the raw data that had been removed in the 2006 version. These errors were manually reintroduced in order to recreate the data set as exact as possible.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

languages, this information is predicted in a two-step process where a finite-state analyzer produces a set of possible annotations for a given verb form, which is then disambiguated by a statistical model trained on gold-standard data (for Czech, see Spoustová et al. [2009]; for Hungarian, see Zsibrita, Vincze, and Farkas [2010]). The German data set was cross-annotated by applying statistical tools[7] trained on the gold standard annotation. Contrary to the Czech and Hungarian data sets, lemma, POS, and morphological information were annotated in three steps, each building upon the preceding one.

In preparation for the experiments, we made two changes to the annotation in the Czech and the Hungarian treebanks in order to allow for a more fine-grained analysis. First, we copied the SubPOS feature value[8] over to the respective POS column (gold to gold, predicted to predicted). This helps us in doing a more fine-grained evaluation, which is based on certain POS tags, but it also allows us to formulate linguistic constraints in the ILP parser more precisely, as we will see in Section 6.1. The German POS tag set is already rather specific. We also changed the object labels (*Obj*) in the Czech data set by combining it with the case value in the gold standard morphology (creating *Obj1-7*). This gives us a more fine-grained object distinction for our analysis and it also separates the case-marked objects from the clausal objects, which do not have a case feature and therefore keep the original *Obj* label.[9]

In order to learn the mapping between case and grammatical functions, the parser relies on the automatically predicted morphological information in the data sets. When the parser is trained on predicted morphology, in principle, it has the chance to adapt to the errors of an automatic morphological analyzer. We will see in Section 5, however, that this does not seem to happen very often. Therefore, if we want the parser to perform well, we need to predict morphological information with high quality. Table 2 shows the prediction quality of the automatic morphological analyzers in the three data sets. On the left-hand side, precision and recall are shown for the phi-features for the whole data set; on the right-hand side, only those words were evaluated where the predicted POS tag matched the gold standard one. We see that Czech and Hungarian achieve high scores on all three features, with Czech achieving over 95% for each feature, and Hungarian over 94% recall and almost 98% precision. In contrast, we find a rather mediocre annotation in the German data set, where only the number feature can be predicted with comparable quality,[10] and gender and case prediction is rather bad. To a certain extent, the lower performance for German compared to Czech can be explained by the more informed annotation tool for Czech. The German data set was annotated by purely statistical tools whereas the Czech annotation tool uses a finite-state lexicon to support the statistical disambiguator.

Hungarian shows a big gap between precision and recall (97.83% and 94.11% for case) when evaluating all words, but the performance on the words with the correct POS tag is almost perfect (99.22% for case!). The reason lies in the POS recognition. The Hungarian POS tag set uses a category *X* as a kind of a catch-all category where annotators would put tokens they could not assign anywhere else. The precision for this class is below 10%, because the tool is classifying a considerable amount of proper nouns (*Np*) as *X*. The class *X*, however, does not get a morphological specification so that about

7 Mate-tools by Bernd Bohnet: `http://code.google.com/p/mate-tools`.
8 The SubPOS feature distinguishes subcategories inside the main POS categories and is part of the morphological description (see Figure 2).
9 Prepositional objects headed by prepositions (pos: RR, RF, RV) were also excluded.
10 There are only two values to predict though.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Table 2
Annotation quality of the phi-features (case, gender, and number) for all words and for those words with a correctly predicted POS tag.

		all		correct POS	
		precision	recall	precision	recall
Czech	case	95.73	95.63	96.06	96.06
	gender	97.59	97.45	98.03	98.03
	number	98.18	98.08	98.47	98.47
German	case	88.69	88.51	89.26	89.06
	gender	90.16	89.99	90.95	90.74
	number	96.18	95.63	96.92	96.61
Hungarian	case	97.83	94.11	99.22	99.22
	number	98.64	95.91	99.88	99.88

3,500 out of 12,500 proper nouns do not receive a case and a number value at all. The reason for the poor morphological annotation in Hungarian is apparently not a problem of an ambiguous morphology, it is simply a problem of the POS recognition. We already know that Hungarian is an agglutinating language. The case paradigm of Hungarian, although comprising about 20 different case values, does not show syncretic forms with the exception of a regular genitive-dative syncretism. Whereas in Hungarian, getting the POS correct effectively means getting case and number correct, the results in Table 2 for Czech and German[11] are not much better for words with correctly predicted POS tags than for all words. In Czech and German, this is a problem of the syncretism in the morphological paradigms.

The low syncretism in the Hungarian case paradigm is due to the agglutinating nature of its morphological system. Because every feature (e.g., case) is signaled by its own morpheme, a syncretism in the system would erase the distinction between the syncretic forms. Because Hungarian uses the same case paradigm for all words, a regular syncretism would mean that a certain distinction can no longer be made in the language.[12] In fusional languages, an inflection morpheme signals more than one feature value. Many syncretisms can thus be disambiguated by the other feature values or by agreement with dependents, as is done in the German noun phrase. We learn two things from these findings: First, we may need different approaches for handling morphology in fusional languages like Czech and German than we do for agglutinating languages like Hungarian. And second, the category *morphologically rich* encompasses

11 The fact that for German, precision and recall differ is due to the independency of the POS tagger and the morphological analyzer. In the German data, the morphological analyzer is not bound to a certain feature template determined by the POS of the word, so that, in principle, it can assign case to verbs and tense to nouns. This is not the case for the Czech and Hungarian analyzers. Precision, recall, and F-score measured over all possible values amount to simple accuracy in those languages.

12 One of the reviewers pointed out to us that Turkish as an agglutinating language also shows much morphological ambiguity. That is correct and this also holds for Hungarian. The case paradigm itself seems to have no syncretism in Turkish, however. The ambiguity rather comes from interaction with vowel harmony and definiteness marking. The syncretism between genitive and dative case in the Hungarian case system is more of a puzzle. Our best guess is that the distribution of these cases is so different that the context can disambiguate them relatively easily.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

languages that are not only different from English but also show important differences among each other that we should take into account when devising parsing technology.

5. Experiment 1

Having examined the quality of the predicted morphological information in the data sets, we can now investigate how the parser deals with this information. We proceed as follows: We train three different models for each language, one using gold standard morphology, one using predicted morphology, and one using no morphological information (henceforth GOLD-M, PRED-M, and NO-M). Comparing the performance of these three models allows us to see the effect that the morphological information has on the parsing performance. The model using gold morphology serves as an upper bound where we can observe the behavior of the parser when it is not disturbed by errors coming from the automatic morphological analyzers. Note that this model is very unrealistic in the sense that syncretisms are fully resolved in the morphological information. The model using predicted morphology serves as a realistic scenario where we can observe the problems introduced by imperfect preprocessing and propagated errors in the pipeline (e.g., due to syncretism). And finally, the model using no morphology shows us how much non-morphological information contributes to the parsing performance. In comparison with the other two models, we can then see the contribution of morphological information[13] to the parsing process. All models use the same predicted lemma and POS information as discussed in the previous section.

5.1 Experimental Set-up

We performed a five-fold cross annotation[14] on the training portions of the data sets of Czech and German, and on the whole subcorpus of Hungarian, varying the morphological annotation as described. The overall parsing performance is shown in Table 3, where the German and the Hungarian scores exclude punctuation and the Czech scores include them.[15]

Table 3 gives us the usual picture that has been noticed in several shared tasks on dependency parsing for multiple languages (e.g., CoNLL-ST 2006, 2007, 2009). The performance on German is pretty high, although not as high as it would be for English, and the performance on Czech is rather low. Note the extreme divergence between labeled (LAS) and unlabeled attachment score (UAS) for Czech.[16] For Hungarian, the performance is comparable to Czech in terms of UAS but the LAS for Hungarian is better. We also see the expected ordering in performance for the models using different kinds of morphological information. The gold models always outperform the models using predicted morphology, which in turn outperform the models using no morphological information. Note, however, that whereas the performance on German does not degrade very much when using no morphological information, it is very

13 It should be noted that by morphological information we always mean the complete annotation available in the treebanks. Although we concentrate in the analysis on the phi-features (gender, number, case), the models using morphological information always use the whole set, including also, for example, verbal morphology.

14 The number of iterations during training was set to 10.

15 Punctuation in the Czech data set is sometimes used as the head in coordination.

16 This is due to the way the Czech data label certain phenomena, which makes it difficult for the parser to decide on the correct label. See Boyd, Dickinson, and Meurers (2008, pages 8–9) for examples.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Table 3
Overall performance of the Bohnet parser on the five-fold cross annotation for every language and different kind of morphological annotation. All results in percent. LAS = labeled attachment score; UAS = unlabeled attachment score. Results for German and Hungarian are without punctuation. Best score for Czech on the CoNLL 2009 Shared Task was by Gesmundo et al. (2009), best score for German was by Bohnet (2009), best score for Hungarian on the CoNLL 2007 Shared Task was by Nivre et al. (2007a). Best CoNLL 09/07 results were obtained on different data sets.

	Czech		German		Hungarian	
	LAS	UAS	LAS	UAS	LAS	UAS
GOLD-M	82.49	88.61	91.26	93.20	86.70	89.70
PRED-M	81.41	88.13	89.61	92.18	84.33	88.02
NO-M	79.00	86.89	89.18	91.97	78.04	86.02
best on CoNLL 09/07	80.38	–	87.48	–	80.27	83.55

harmful for Hungarian to do so (78.04% LAS for NO-M in comparison with 84.33% LAS for PRED-M). The Czech results lie in between. To give a general impression of the performance of the parser, the last row shows parsing results for the three languages reported in the literature. The results have been obtained on different data sets, however, so a direct comparison would be invalid.

5.2 Analysis

Although the scores in Table 3 reflect the quality of the parser on the complete test data, we would not expect case morphology to influence all of the functions. We will therefore go into more detail and concentrate on nominal elements (nouns, pronouns, adjectives, etc.)[17] and core grammatical functions (subjects, objects, nominal predicates, etc.) because in our three languages, nominal elements carry case morphology to mark their syntactic function. Core grammatical functions are vital to the interpretation of a sentence because they mark the participants of a situation. We exclude clausal and prepositional arguments, which can fill the argument slot of a verb but would not be marked by case morphology. Table 4 shows the encoding of the core grammatical functions in the three treebanks.

Table 5 shows the performance of the parsing models for each of the three languages on the core grammatical functions. As described in Section 4, we split the object function for Czech according to its associated case value. The results are shown for each of the three models with GOLD-M on the left, PRED-M in the middle, and NO-M on the right.

The results shown for the NO-M models indicate again that morphology plays a bigger role in Czech and Hungarian for determining the core grammatical functions than it does for German. The performance on all grammatical functions except the rather rare genitive object is generally higher for German, showing that to a large

17 We determine a nominal element by its gold standard POS tag:
Czech: AA, AG, AM, AU, C?, Ca, Cd, Ch, Cl, Cn, Cr, Cw, Cy, NN, P1, P4, P5, P6, P7, P8, P9, PD, PE, PH, PJ, PK, PL, PP, PQ, PS, PW, PZ.
German: ADJA, ART, NE, NN, PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS, PRF, PWS, PWAT.
Hungarian: Oe, Oi, Md, Py, Oh, Ps, On, Px, Pq, Mf, Pp, Pg, Mo, Pi, Pr, Pd, Mc, Np, Af, Nc.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Table 4
Core argument functions and their encoding in the different treebanks. The different object labels for Czech have been introduced by us. The original function is simply Obj.

	PDT 2 (Czech)	TiGer (German)	HunDep (Hungarian)
subject	Sb	SB	SUBJ
nominal predicate	Pnom	PD	PRED
object	Obj1-7	OA, OA2, DA, OG	OBJ, DAT

Table 5
Precision, recall, and F-score (LAS) for core grammatical functions marked by case. We omit locative objects in Czech, and second accusative objects in German, due to their low frequency.

		GOLD-M			PRED-M			NO-M		
Czech	freq	prec	rec	f	prec	rec	f	prec	rec	f
subject	38,742	89.29	91.18	90.22	83.96	87.01	85.46	74.10	78.82	76.39
obj (acc)	21,137	92.50	93.35	92.93	85.25	83.01	84.12	73.42	72.02	72.71
predicate	6,478	89.07	87.14	88.09	88.24	86.00	87.11	82.34	78.19	80.21
obj (dat)	3,896	83.18	85.68	84.41	80.21	78.88	79.54	74.29	48.05	58.35
obj (instr)	1,579	71.38	66.50	68.85	67.74	62.51	65.02	58.93	35.53	44.33
obj (gen)	1,053	86.69	77.30	81.73	80.42	62.39	70.26	74.60	48.81	59.01
obj (nom)	167	57.63	40.72	47.72	56.97	29.34	38.74	48.67	32.93	39.29

		GOLD-M			PRED-M			NO-M		
German	freq	prec	rec	f	prec	rec	f	prec	rec	f
subject	45,670	95.11	96.05	95.58	89.95	91.23	90.59	88.32	89.86	89.08
obj (acc)	23,830	93.93	94.80	94.36	84.83	84.89	84.86	82.20	83.35	82.77
obj (dat)	3,864	89.56	87.73	88.64	79.17	64.44	71.05	77.09	50.78	61.23
predicate	2,732	78.07	73.35	75.64	75.80	72.91	74.33	76.20	71.01	73.51
obj (gen)	155	80.25	41.93	55.08	60.66	23.87	34.26	52.94	17.42	26.21

		GOLD-M			PRED-M			NO-M		
Hungarian	freq	prec	rec	f	prec	rec	f	prec	rec	f
subject	11,816	88.34	91.57	89.93	84.96	88.15	86.53	64.58	66.44	65.50
obj (acc)	9,326	93.63	94.22	93.92	92.36	92.70	92.53	66.23	63.86	65.03
obj (dat)	1,254	80.55	76.95	78.71	75.57	71.53	73.49	58.36	30.62	40.17
predicate	941	81.05	75.45	78.15	77.39	72.37	74.79	72.49	71.41	71.95

extent the parser is able to use information from lexicalization and configurational information (Seeker and Kuhn 2011). Results for Czech and Hungarian are lower in the NO-M models. They improve by large margins when switching to predicted morphology. Czech accusative objects improve from 72.71% F-score to 84.12% F-score in the PRED-M model. In Hungarian, the F-scores for dative objects improve by over 33 percentage points to 73.49% F-score when switching to the PRED-M model. In contrast, although all the scores improve for German, improvements are generally low when switching from the NO-M to the PRED-M model. The biggest improvement happens

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

for dative objects, which increase by about 10 percentage points, but for subjects, the improvement is just over one percentage point. This is in line with the general idea that German is a borderline case between morphologically poor configurational languages like English and morphologically rich non-configurational languages like Czech or Hungarian. We already saw this general trend in Table 3, but the effect is much larger if we consider those functions that are directly marked by morphological means in the language.

If we now turn to the GOLD-M models, we see that in general, German and Czech benefit more from the gold standard morphological annotation than Hungarian. Knowing that Hungarian does not have much form syncretism in its inflectional paradigms, this is not really surprising. There is, however, still a gain of information because the effect of the wrong POS tags in Hungarian is eliminated in the GOLD-M model. An effect that comes out very clearly is the improvement for subjects and accusative objects for Czech and German when moving from predicted to gold morphology, because the typical syncretism between nominative and accusative in the neuter gender in Indo-European languages (cf. Table 1) is correctly disambiguated: Comparing the performance on subjects (marked by nominative case) and accusative objects, we see a considerable improvement between 5 percentage points for Czech subjects and almost 10 percentage points for German accusative objects when switching to gold morphology. This improvement does not happen for Hungarian, where there is no such syncretism. The gold morphology acts as an oracle here and circumvents the ambiguity problem that a pipeline approach to predicting morphological information prior to parsing has.

Another interesting observation related to the way the parser works is that for all languages, predictions are less accurate for the less frequent functions. The general order for all three languages from most frequent to least frequent is *subjects* > *accusative objects* > *predicates/dative objects* > *instrumental/genitive objects*. For all languages, the parser's quality of annotation follows this ordering. This effect comes from the statistical nature of the parsing system, which will in case of doubt resort to the more frequent function. A clear sign is that for rare objects, the precision is always higher than the recall. As an example, notice the performance of the parsing models on dative and genitive objects. The parser annotates genitive objects if it has strong evidence, hence the high precision, but it frequently fails to find it in the first place, hence the low recall. Because the NO-M models do not have morphological information, they can only rely on lexicalization and contextual information to determine the correct grammatical function. We can see this ranking in all the models regardless of the amount of morphological information available, although the differences are much smaller for the more informed models.

Finally, we see that the benefit from morphological information is comparatively low for nominal predicates. It seems that the non-morphological context already provides much useful information (e.g., the copular verbs).

5.3 Analysis of Confusion Errors

We now ask ourselves if the parser utilizes the morphological information, in our case the case morphology, correctly. In principle, there are two possible scenarios: (1) the feature model of the parser does not integrate the morphological annotation in a useful way, so that the parser has difficulties learning the association between case values and the grammatical functions; (2) There is nothing wrong with the feature model, but the

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

morphological annotation is not good enough and causes problems because the parser gets incorrect information in the features.

To answer this question, we examine the confusion errors made by the parser. If the parser uses morphological information correctly, we expect it to confuse labels that can all be signaled by the same case value. For example, if the parser learns the association between nominative and subject/predicate properly, we would still expect it to make errors in confusing these two functions. Because the mapping between case and grammatical function is one-to-many, knowing the case value reduces the number of possible functions but the final decision between these functions must be made by non-morphological information. The effect should be strongest in the GOLD-M models because the morphological information is correctly disambiguated. Consequently, we expect the same results for the PRED-M models blurred by additional errors introduced by an imperfect morphological prediction. If, however, the parser does not learn the mapping or has no access to morphological information, we expect confusion errors all across the case paradigms.

To start with the last hypothesis, we examine the confusion errors with subjects made by the parser using the NO-M models. Subjects are marked by nominative case in all three languages, with Czech allowing for dative and genitive subjects under special circumstances. The NO-M models do not have access to morphological information and should therefore mix up functions regardless of the case value that would usually distinguish them. Table 6 shows the top five confusion errors made by the NO-M models on the subject function. The values are split for correct and incorrect head selection to tease apart simple label classification errors from errors involving label classification and attachment.

Table 6
Top five functions with which subjects were confused when parsing with the NO-M models. _M marks a coordinated function in Czech.

	Czech					**German**			
NO-M	correct head		wrong head		NO-M	correct head		wrong head	
rank	label	freq	label	freq	rank	label	freq	label	freq
1	Obj4	4,996	Atr	2,644	1	OA	2,680	OA	1,498
2	Pnom	1,261	Obj4	981	2	PD	776	NK	906
3	Adv	811	Sb_M	948	3	DA	458	DA	431
4	Obj3	752	Adv	273	4	EP	301	AG	313
5	Obj7	380	Obj_M	245	5	MO	219	CJ	296

	Hungarian			
NO-M	correct head		wrong head	
rank	label	freq	label	freq
1	OBL	3,029	ATT	1,116
2	OBJ	1,505	Exd	574
3	PRED	250	COORD	313
4	ATT	185	OBL	311
5	DAT	152	OBJ	139

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

The results in Table 6 confirm the expectation that confusion errors appear regardless of the case value involved, which is no surprise given that the models do not have access to morphological information: For Czech, when the head was chosen correctly, *Obj4, Obj3*, and *Obj7* (accusative, dative, and instrumental objects, respectively) are all signaled by a different case value and their confusion rates follow their frequency in the data. *Pnom* (nominal predicates) are expected because they are also signaled by nominative case as are subjects. If the head was chosen incorrectly, the parser assigns *Obj4* and coordinated subjects and objects (*Sb_M, Obj_M*). Adverbial (*Adv*) and attributive functions (*Atr*) are expected as they mark adjunct functions that can be filled by nominal elements. For German, we see confusions with the object functions (accusative *OA* and dative objects *DA*), predicates (*PD*), and the *EP* function marking expletive pronouns in subject position. Both are marked by nominative case. Furthermore, the parser makes confusion errors with *MO, NK*, and *AG*, which are the three adjunct functions that can be filled by nominal elements (e.g., *AG* marks genitive adjuncts). *CJ* finally marks coordinated elements, which is an expected error if the head was chosen incorrectly, but, unlike in the Czech treebank, we cannot tell by the coordination label the particular function the element would have if it were not coordinated. In Hungarian, we also have errors across the board, with argument functions not marked by nominative case (accusative objects *OBJ*, dative objects *DAT*), the predicate function *PRED*, and all types of adjuncts (*ATT* [attributives] and *OBL* [obliques]). Obliques are especially interesting in Hungarian because the language has only a small number of prepositions. Most oblique adjunct functions are realized by a particular case (hence the about 20 different case values), which for a parsing model using no morphological information makes it rather difficult to distinguish them from the core argument functions. In summary, we find the expected picture of confusion errors across the case paradigms.

Turning now to the GOLD-M models, we can test whether the parser is able to learn the mapping between case and its associated functions. If so, we expect confusion errors with functions that are all compatible with the case value of the correct function. Table 7 shows the top five confusion errors that the GOLD-M models made on the subject function. Here, we see a completely different picture compared with the NO-M model errors in Table 6. In all three languages, we find—regardless if the head is correct or not—confusions only with functions that are compatible with the nominative case. In Czech, subjects are mostly confused with predicates (*Pnom*) and coordinated subjects (*Sb_M*). *ExD* marks suspended nodes moved because of an elliptical constructions. The label does not tell whether the node would be a subject with regard to the empty node but it may be, so it is compatible with nominative case. *Atr* between nominal elements may mark close appositions like the one in Figure 1, which would be marked as nominative by default. *ObjX* marks objects with no annotated case value (mostly for foreign words). Of all the functions, only *Obj4* cannot be signaled by nominative case. If one checks those 69 cases, only 22 are annotated with accusative case in the gold standard, the rest consist mostly of various, high-frequent numerals in neuter gender and quantifiers, most of which are ambiguous between nominative and accusative. In these cases, lexicalization seems to overrule the case feature. We get the same picture for German and Hungarian, both models making errors that are compatible with the nominative case value. Of the 112 errors with accusative objects (*OA*) in German, only 36 have the correct case value in the gold standard. Unlike in the Czech and the Hungarian treebank, the morphological annotation in TiGer contains a considerable number of errors. We then conclude that for subjects, the parser indeed has no problem learning that subjects are marked by nominative case.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Table 7
Top five functions with which subjects were confused when parsing with the GOLD-M models. _M marks a coordinated function in Czech.

Czech					German				
GOLD-M	correct head		wrong head		GOLD-M	correct head		wrong head	
rank	label	freq	label	freq	rank	label	freq	label	freq
1	Pnom	583	Sb_M	1,142	1	PD	773	NK	555
2	ObjX	102	Atr	711	2	EP	323	CJ	245
3	Adv	102	ExD_M	162	3	MO	117	PNC	139
4	Obj4	69	ExD	145	4	OA	112	PD	129
5	ExD	45	Pnom	65	5	PH	96	APP	127

Hungarian				
GOLD-M	correct head		wrong head	
rank	label	freq	label	freq
1	PRED	264	ATT	678
2	Exd	102	Exd	494
3	OBL	94	COORD	249
4	ATT	90	NE	32
5	OBJ	50	DET	22

Next, we examine the accusative objects and compare the performance of the GOLD-M models with their respective PRED-M counterparts to assess the effect of predicted morphological information. Table 8 shows the confusion errors for the accusative objects. On the left, the GOLD-M errors are shown; on the right we see the PRED-M errors. For the GOLD-M models, the picture is basically the same as with the subjects, with the small exception that all three languages show confusion with subjects under the top five.[18] Although the effect is not strong, it shows that the statistical model can sometimes overrule the morphological features even for the gold standard morphology.

The most interesting effect, however, happens when switching to predicted morphological information. The overall number of errors increases, but the biggest increase occurs for subjects in German (*SB*) and in Czech (*Sb*), although the same is not observable in Hungarian (*SUBJ*). Of the 2,945 confusion errors in Czech, where the PRED-M model incorrectly predicts an accusative object, 891 have been classified as accusative despite being nominative in the gold standard and 1,505 have been classified as nominative although being accusative. If we check the gender of these instances, we find the overwhelming majority to be neuter, feminine, or masculine inanimate, exactly those genders whose inflection paradigms show syncretism between nominative and accusative forms. We find the same effect in the German errors. The syncretism in the two languages causes the automatic morphological analyzers to confuse these case

18 The *AuxT* label in the Czech errors is used to mark certain kinds of reflexive pronouns, which can be in accusative or dative case. The criterion for deciding whether a reflexive pronoun is labeled *AuxT* or *Obj4* (i.e., accusative object) is whether the governing verb denotes a conscious or unconscious action. This is a very tough criterion to learn for a dependency parser. In any case, however, *AuxT* is perfectly compatible with accusative case.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Table 8
Top five functions with which accusative objects were confused when parsing with the gold (left) and predicted (right) morphology models. _M marks a coordinated function in Czech.

Czech

GOLD-M	correct head		wrong head		PRED-M	correct head		wrong head	
rank	label	freq	label	freq	rank	label	freq	label	freq
1	Adv	274	Obj_M	750	1	Sb	2,354	Atr	687
2	AuxT	270	Atr	172	2	Adv	262	Obj_M	660
3	Sb	69	ExD_M	67	3	AuxT	256	Sb	594
4	ExD	34	Adv	65	4	Obj3	137	Sb_M	108
5	AuxR	28	Atv	53	5	Obj2	109	ExD_M	94

German

GOLD-M	correct head		wrong head		PRED-M	correct head		wrong head	
rank	label	freq	label	freq	rank	label	freq	label	freq
1	MO	283	NK	357	1	SB	2,176	SB	1,329
2	SB	112	CJ	191	2	DA	610	NK	606
3	DA	55	SB	121	3	MO	308	CJ	365
4	CJ	43	APP	97	4	CJ	46	AG	137
5	EP	25	MO	55	5	EP	40	APP	136

Hungarian

GOLD-M	correct head		wrong head		PRED-M	correct head		wrong head	
rank	label	freq	label	freq	rank	label	freq	label	freq
1	OBL	90	COORD	119	1	OBL	119	COORD	140
2	ATT	60	Exd	81	2	SUBJ	86	Exd	111
3	SUBJ	50	ATT	44	3	ATT	65	ATT	78
4	Exd	23	OBL	18	4	Exd	19	ROOT	18
5	MODE	14	ROOT	13	5	MODE	16	OBL	15

values more often, which subsequently leads to errors in the parser due to the pipeline architecture. That the parser so frequently falls for incorrect annotation is more proof that it has learned the mapping between case and its associated grammatical functions. As expected, we do not find this effect for Hungarian. As we discussed in Section 4, there is almost no syncretism in the Hungarian case paradigm, which therefore does not lead to this kind of error propagation. The slight increase in errors in Hungarian is instead related to the POS errors and their influence on missing morphological information than the quality of the predicted morphology itself.

For reasons of space and because it would not contribute anything new to the picture, we will not go into detail for the errors for the remaining grammatical functions. We conclude that learning the morphological dependencies that hold for a language (cf. the four types by Nichols [1986]) can be facilitated by a statistical model. When presented with gold standard morphological information, the parser performance improves considerably over the model without morphological information for all three

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

languages. The error analysis shows that the parser learns the mapping between case and grammatical function, which also shows that the feature model of the parser integrates the information in a useful way. In the more realistic scenario using predicted morphology, however, the parser starts making more mistakes for Czech and German that are caused by errors of the automatic morphological predictors, which are propagated through the pipeline model. This effect does not occur for Hungarian. The syncretism in the inflectional paradigms in Czech and German makes the task of learning the morpho-syntactic rules of a language much more difficult for a statistical parser in a pipeline architecture. With a high amount of syncretism, it is simply not sensible to fully disambiguate certain morphological properties of a word (e.g., case) without taking the syntactic context into account.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

6. Case as a Filter

From Experiment 1 we learned that one of the problems when parsing morphologically rich languages like Czech and German is the propagation of annotation errors in the processing pipeline and the unreliable morphological information. The problem is that the parser learns a mapping between case values and grammatical functions but the predicted morphology delivers the wrong case value. As a solution to this problem, Lee, Naradowsky, and Smith (2011) have proposed a joint architecture where the morphological information is predicted simultaneously with the syntactic structure, so that both processes can inform and influence each other. This puts morphological prediction and syntactic analysis on the same level. We choose a different approach here: We keep the basic pipeline architecture, because it works very efficiently. We support the parser, however, with constraints that model the possibly underspecified morphological restrictions grounded in the surface forms of the words. Especially for the core argument functions, a morphological feature like case first and foremost serves as a morpho-syntactic means to support the syntactic analysis by overtly marking syntactic relations and thus reducing the choice for the parser. For example, if a word form morphologically cannot be accusative, the parser should not consider grammatical functions that are signaled by accusative in the language. Case acts here as a filter on the available functions for the morphologically marked element. Interpreting the role of case as a filter, we can use the case feature as a formal device to restrict the search space of the parser. This is different from the joint model, where morphology and syntax are predicted at the same time, because the parser will not fully disambiguate a token with respect to its morphology if the syntactic context does not provide the necessary information. Another thing that we learned from the first experiment is that although the predicted morphology is not completely reliable, it is still much better than using none at all, especially for Czech and Hungarian (see difference between PRED-M and NO-M models in Table 5). In the following, we will therefore still use the predicted morphology as features in the statistical model in combination with the filter. In this architecture, the parser gets statistical information from the feature model to prefer a particular analysis, but the constraints will block this option if it does not comply with the morphological specification of the words. The parser then needs to choose a different option.

In order to implement the constrained parser, we use a parsing approach by Martins, Smith, and Xing (2009) using integer linear programming. It is related to the Bohnet parser in the sense that it is also a graph-based approach, but it allows us to elegantly augment the basic decoder with linguistically motivated constraints

(Klenner 2007; Seeker et al. 2010). ILP is a mathematical tool for optimizing linear functions and was first used in dependency parsing by Riedel and Clarke (2006), who performed experiments on Dutch using linguistically motivated constraints as we will do. Martins, Smith, and Xing improved the formulation considerably so that the parser would output well-formed dependency trees without the need for iterative solving. In our ILP parser, we use the formulation by Martins, Smith, and Xing extended to labeled dependency parsing. Like the Bohnet parser, the ILP parser consists of a decoder and a statistical feature model. Whereas the feature model remains basically the same, the decoder is implemented using ILP. The formulation represents every possible arc that might appear in the parse tree as a binary variable (arc indicator), where 1 signals the presence of the arc in the tree, and 0 signals its absence (see also Figure 3). Each such arc indicator variable is weighted by a score assigned by the statistical model that is learned from a treebank. During decoding,[19] the parser searches for the highest scoring combination of arcs that also fulfills the global tree constraints as well as any other global constraints that may be added to the equations to model, for instance, linguistic knowledge. The tree constraints ensure that every word in the tree has exactly one head and that there are no cycles in the tree. Martins, Smith, and Xing use the single commodity flow formulation by Magnanti and Wolsey (1995) to enforce the tree structure. The idea is that the root node sends N units of flow through the tree (with N being the number of words in the sentence) and every node in the tree consumes one unit. If every node consumes exactly one unit of flow and every node can have only one parent node, then the tree must be connected and acyclic.

$$\max \sum_{h \in H} \sum_{d \in N} \sum_{l \in L} \omega^{l}_{dh} a^{l}_{dh} \tag{1}$$

$$\sum_{h \in H} \sum_{l \in L} a^{l}_{dh} = 1 \quad \forall d \in N \tag{2}$$

$$|N| \sum_{l \in L} a^{l}_{dh} \geq f_{dh} \quad \forall d \in N, \forall h \in H \tag{3}$$

$$\sum_{h \in H} f_{dh} - \sum_{g \in N} f_{gd} = 1 \quad \forall d \in N \tag{4}$$

$$\sum_{d \in N} f_{dRoot} = |N| \tag{5}$$

$$a \in \{0, 1\}, f \in \mathbb{Z} \tag{6}$$

Let N be the set of words in a sentence, $H = N \cup \{Root\}$ is the set of words plus an artificial root node, and L is the set of function labels. For every sentence, Equations (1)–(6) constitute the equation system that the constraint solver has to solve in order to find the highest scoring dependency tree. Equation (1) shows the objective function, which is simply the sum over all binary arc indicator variables $a \in A = N \times H \times L$ weighted by their respective score ω. Equation (2) restricts for every dependent d the number of incoming arcs to exactly one. It thus makes sure that every word will end up with

19 We use the GUROBI constraint solver: `www.gurobi.com`, version 4.0.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

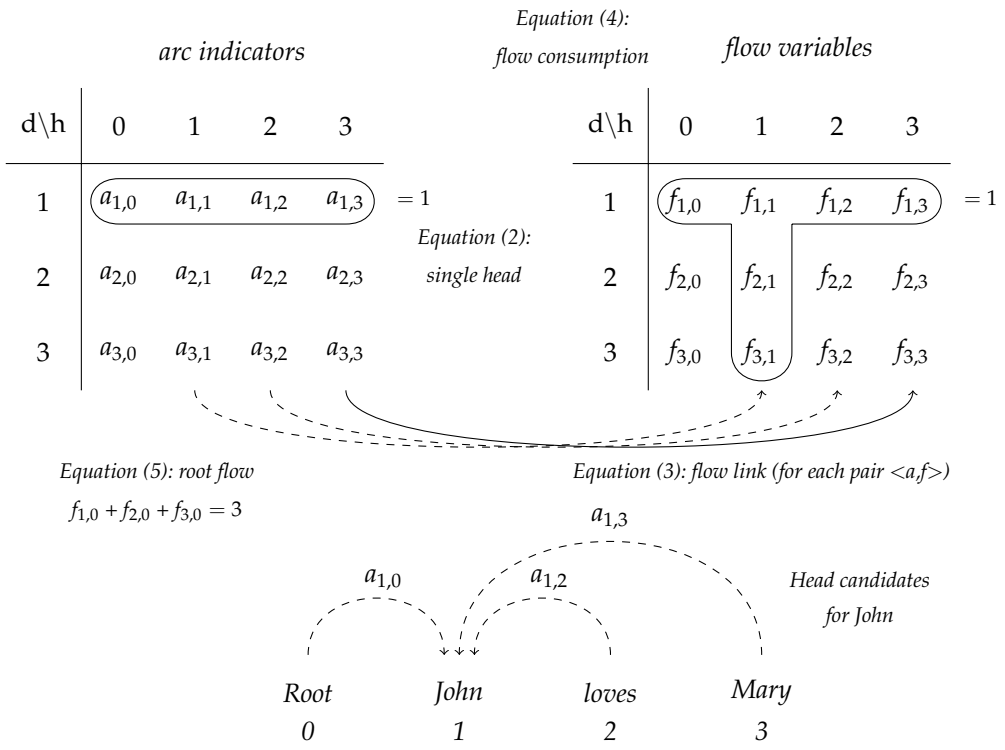

Figure 3
Schematic description of the unlabeled first-order model for the example sentence *John loves Mary*. The constraints are shown for the dependent (d) *John*. There are three head (h) candidates, from which the decoder needs to choose one because of the single head constraint (Equation (2)). Equations (3), (4), and (5) show as an example how the flow constraints are applied to ensure a tree structure. Equation (3) links each arc indicator to one flow variable making sure that only active arcs (those that are set to 1) carry flow > 0. Equation (5) sends three units of flow from the root, one for each other token in the tree. Equation (4) finally forces the flow difference between the incoming arc (horizontal part) of each node (except root) and the flow on all outgoing arcs (vertical part) to be exactly 1, thus making sure that each node consumes one unit of flow. To find the optimal tree, the sum over the weights of all arc indicators that are set to one is maximized.

exactly one head. Equations (3)–(5) model the single commodity flow. A set of integer variables $F = N \times H$ is introduced to represent the flow on each arc. Equation (3) links every flow variable that represents the flow between two nodes to the set of arc indicator variables that can connect these two nodes. If there is no arc between the two nodes (all indicator variables are 0), the flow must be 0 as well. If one arc indicator is 1, then the flow variable can take any integer value between 0 and $|N|$. Equation (4) enforces the consumption of one unit of flow at each node by requiring the difference between incoming and outgoing arcs to be exactly one. Equation (5) finally sets the amount of flow that is sent by the artificial root node to the number of words in the sentence. Note that this does not force the tree structure to be single-rooted, because the artificial root node can have multiple dependents. It can be done by an additional constraint that sets the number of dependents for the root node to one. Figure 3 shows an example for the basic formulation.

Martins, Smith, and Xing (2009) propose several extensions to the basic model; for example, second-order features, which introduce new variables for each combination

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

of two arc indicator variables into the ILP model. For our parser, we implemented the second-order features that they call *all grandchildren* and *all siblings*. They also state that the use of second-order features in the decoder renders exact decoding intractable, and they propose several techniques to reduce the complexity, which we also apply to our parser: (1) Before parsing, the trees are pruned by choosing for each token the ten most probable heads using a linear classifier that is not restricted by structural requirements, and (2) The integer constraint is dropped, such that the variables can now take values between 0 and 1 instead of either 0 or 1. The dropping of the integer constraint can lead to inexact solutions with fractional values. To arrive at a well-formed dependency tree, we then use the first-order model in Equations (1)–(6) to get the maximum spanning tree, this time using the fractional values from the actual solution as arc weights. Two other techniques that we apply are related to the arc labels: (1) We use an arc filter (Johansson and Nugues 2008) like the Bohnet parser, which blocks edges that did not appear in the training data based on the POS tags of the dependent and the head, and the label, and (2) We do not include labels in the second-order variables.

The feature set of the ILP parser is similar to but not identical to one in the Bohnet parser. The ILP parser uses loss-augmented MIRA for training (Taskar et al. 2005), which is similar to the MIRA used in the Bohnet parser. We set the number of training iterations to 10 as well.

6.1 Morpho-Syntax as Constraints

Using case as a filter for the decoder requires an underspecified symbolic representation of morphological information that we can use to define constraints. This allows us to have an exact representation of syncretism controlling the search space of the parser. The case features of a word are represented in the ILP decoder as a set of binary variables M for which 1 signals the presence of a particular value and 0 signals its absence. For Hungarian, we only model the different case values, which leads to one binary variable for each of the values. For Czech and German, we also include the gender and the number features which then gives, for each case marked word, a binary variable for every combination of the case, number, and gender values. The values of the morphological indicator variables are specified by annotating the data sets with underspecified morphological descriptions that are obtained from finite-state morphological analyzers.[20] If a certain feature value is excluded by the analyzers, the value of the indicator variable for this feature is fixed at 0, which then means that the decoder cannot set it to 1. This way, all morphological values that cannot be marked by the form of the token (according to the morphological analyzer) are blocked and thereby also all parser solutions that depend on them. Words unknown to the analyzers are left completely underspecified so that each of the possible values is allowed (none of the variables are fixed at 0). The symbolic, grammar-based pre-annotations thus set some of the morphological indicator variables to 0 where the word form gives enough information while leaving other variables open to be set by the parser, which can use syntactic context to make a more informed decision.

We now present three types of constraints that model the morpho-syntactic interactions in the three languages. Their purpose is to help the parser during decoding

20 Czech: `http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Morphology/index.html`; German: Schiller (1994); Hungarian: Trón et al. (2006).

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

to find a linguistically plausible solution. They are inspired by the types of morpho-syntactic interaction that Nichols (1986) describes and guide the parser by enforcing them globally in the final structure. It is important to emphasize that these constraints do not interact with or influence the statistical feature model of the parser. They are applied during decoding when the parser is searching for the highest-scoring tree and prevent solutions that violate the constraints.

The first type of constraints that we apply explicitly formulates the mapping between a function label and the case value that it requires. Equation (7) shows an example of a case licensing constraint for the *DAT* label in Hungarian. A dependent d cannot be attached to a head with label *DAT* if its morphological indicator variable for dative case (m_d^{dat}) is zero.

$$\forall d : \sum_{h \in H} a_{dh}^{DAT} \leq m_d^{dat} \tag{7}$$

The second type of constraint models the morphological agreement between dependents and their heads in noun phrases (Equations (8)–(9)), for instance, determiners and adjectives with their head noun in the noun phrases in Czech and German. In the treebanks, the relation is marked by *NK* for German and *Atr* for Czech.[21] The constraints set the morphological indicators for an adjective and a noun in the following relation: As long as there is no arc (a_{dh}^{NK} is 0) between the adjective (d) and the noun (h), the two constraints allow for any value in the morphological indicator variables of both words. If the arc is established (a_{dh}^{NK} is set to 1), the two constraints form an equivalence forcing all the morphological indicators to agree on their value (i.e., to be both 1 or both 0). We additionally require every word to have at least one morphological indicator variable set to 1. Thus, if there is no solution to the equivalence the arc between the adjective and the noun cannot be established with this function label.

$$m_h^{dat-pl-fem} \leq m_d^{dat-pl-fem} + 1 - a_{dh}^{NK} \tag{8}$$

$$m_h^{dat-pl-fem} \geq m_d^{dat-pl-fem} - 1 + a_{dh}^{NK} \tag{9}$$

For the third type, Equation (10) shows a constraint that was already proposed by Riedel and Clarke (2006). It models label uniqueness by forcing label l to appear at most once on all the dependents of a head (h). Due to the design of the decoder following Carreras (2007), the Bohnet parser has no means of making sure that a particular function label is annotated at most once per head. Table 9 shows the number of times a grammatical function occurs more than once per head in the treebank (TRBK) and how often it was annotated by the models in the previous experiment. Although doubly annotated argument functions almost never appear in the treebank, the parser

21 In German and mostly also in Czech, if an adjective is attached to a noun by *NK* (or *Atr*), they stand in an agreement relation. This fortunate circumstance allows us to bind the agreement constraint to these function labels (and to the involved POS tags). In a (very) small number of cases in the Czech treebank, however, an adjective is attached to a noun by *Atr* but there is no agreement. This happens, for example, if the adjective is actually the head of another noun phrase that stands in attributive relation (*Atr*) to the noun. The *Atr* label was not meant to mark agreement relations, it just happens to coincide for most of the cases. But it might be worth considering whether morpho-syntactic relations like agreement should be represented explicitly in syntactic treebanks.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

frequently annotates them because it has no way of checking whether the function has already been annotated (see also Khmylko, Foth, and Menzel [2009]).

$$\forall h \forall l : \sum_{d \in N} a^{l}_{dh} \leq 1 \tag{10}$$

The global constraint in Equation (10) allows us to restrict the number of argument functions and thus implements a very conservative version of subcategorization frame with which we do not risk coverage problems caused by too restrictive verb frames. For each language, we automatically counted the number of times a function label occurred on the direct dependents of each node in the treebank. Labels that occurred more than once per head with a very low frequency were still counted as appearing at most once if our linguistic intuition would predict that (see, e.g., German subjects in Table 9). For each function label l in these lists, the constraint in Equation (10) was applied.

Table 9
Number of times a core grammatical function was annotated more than once in the treebank (TRBK) by the model using gold morphology (GOLD-M), and by the model using predicted morphology (PRED-M).

	Czech			German			Hungarian		
	TRBK	GOLD-M	PRED-M	TRBK	GOLD-M	PRED-M	TRBK	GOLD-M	PRED-M
subjects	0	772	1,723	44	1,170	2,403	0	586	670
predicates	7	174	190	6	92	108	1	17	19
obj (dat.)	0	28	46	0	33	46	0	9	5
obj (acc.)	22	284	602	2	364	912	0	182	189

Each individual constraint already reduces the choices that the parser has available for the syntactic structure. They exclude additional incorrect analyses, however, by interaction. Figure 4 illustrates the interaction between the three constraints for the German sentence *den Mädchen helfen Frauen* meaning *women help the girls*. Each individual word displays a high degree of syncretism. But when the syntactic structure is decided, many options mutually exclude each other. Constraints (8) and (9) disambiguate *den Mädchen* for dative plural feminine. The case licensing (Constraint (7)) then restricts the labels for *Mädchen* to dative object (*DA*), and Constraint (10) ensures uniqueness by restricting the choice for *Frauen*. The parser now has to decide whether *Frauen* is subject, accusative object, or something else completely. The constraints are applied online during the decoding process. If the statistical model would strongly prefer *Mädchen* to be accusative object, the parser could label it with *OA*. In that case, however, it would not be able to establish the *NK* label between *den* and *Mädchen*, because the agreement constraint would be violated. So, the constraints filter out incorrect solutions but the decoder is still driven by the statistical model.

6.2 Experiment 2

In the second experiment, we now apply the ILP parser to the same data sets that we used in the first experiment, again with a five-fold cross-annotation. We trained two

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

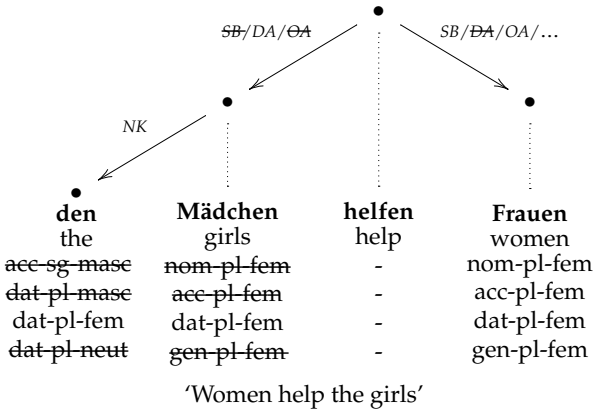

Figure 4
Constraint interaction for the German sentence *den Mädchen helfen Frauen* meaning *women help the girls*.

Table 10
Overall performance of the Bohnet parser and the ILP parser on the five-fold cross annotation for every language. All results in percent. LAS = labeled attachment score, UAS = unlabeled attachment score. Results for German and Hungarian are without punctuation.

	Czech		German		Hungarian	
model	LAS	UAS	LAS	UAS	LAS	UAS
GOLD-M	82.49	88.61	91.26	93.20	86.70	89.70
PRED-M	81.41	88.13	89.61	92.18	84.33	88.02
NO-M	79.00	86.89	89.18	91.97	78.04	86.02
ILP NO-C	81.69	88.09	89.30	91.98	84.01	87.12
ILP C	81.91	88.18	89.93	92.25	84.35	87.39

models for each language, one using the constraints (*c*) and one without the constraints (*no-c*). In both cases, we used the predicted morphology in the feature set. Table 10 shows the parsing results for the ILP parsing models in terms of LAS and UAS in comparison to the results of the Bohnet parser (repeated from Table 3). Both ILP models should be compared to the PRED-M model because they have the most similar feature sets. As can be seen from the results, the ILP parser without constraints performs overall slightly worse than the Bohnet parser and the ILP parser using constraints performs overall slightly better or equal. This shows that both parsers perform on a similar level. The differences between the Czech and the German models (ILP C vs. PRED-M) are statistically significant.[22] The interesting results, however, occur for the argument functions.

Table 11 shows the performance of the unconstrained (*no-c*) and constrained (*c*) ILP models and the PRED-M models of the Bohnet parser on the argument functions. Again,

22 According to a two-tailed t-test for related samples with $\alpha = 0.05$.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Table 11
Parsing results for the unconstrained (NO-C) and the constrained (C) ILP models, and the Bohnet parser in terms of F-score (LAS) for core grammatical functions marked by case. We omit locative objects in Czech, and second accusative objects in German because of their extremely low frequency. ∗ Statistically significant when comparing the performance on a grammatical function for the C model to the PRED-M model ($\alpha = 0.05$, two-tailed t-test for related samples).

	Czech			German			Hungarian		
	NO-C	C	PRED-M	NO-C	C	PRED-M	NO-C	C	PRED-M
subject	85.41	**87.23***	85.46	90.02	**92.91***	90.59	85.05	**87.67***	86.53
predicate	87.13	**90.09***	87.11	72.86	**80.70***	74.33	74.16	**78.88***	74.79
obj (nom)	47.48	**53.19***	38.74	–	–	–	–	–	–
obj (gen)	70.15	**72.54**	70.27	31.41	**42.98**	34.26	–	–	–
obj (dat)	79.99	**80.42**	79.54	65.21	**77.78***	71.05	75.33	**77.92***	73.49
obj (acc)	84.27	**86.79***	84.12	83.74	**87.96***	84.86	91.96	**93.21***	92.53
obj (instr)	67.36	**68.76**	65.02	–	–	–	–	–	–
all arg funcs	84.33	**86.37***	84.21	86.27	**90.11***	87.24	86.87	**89.04***	87.78
all other	81.37	**81.37**	81.05	89.79	89.88	**89.98**	82.73	82.86	**83.43**

we only evaluated those tokens that actually carry case morphology, as we did in the first experiment. For each language, the best results are in boldface. In addition to the results for the different argument functions, a total score is computed over all argument functions (all arg funcs) and another is computed over all tokens that are not included in the first score (all other). The latter illustrates the performance of the parsing models on the functions that are not marked by case morphology.

For each language, we get the same basic picture: Although the unconstrained ILP model performs slightly worse than (German, Hungarian) or equally well as (Czech) the PRED-M model of the Bohnet parser, the constrained ILP model clearly outperforms both on the argument functions. On each of them, the constrained ILP model improves over the other two models, raising the score by 1 percentage point for (for example) subjects in Hungarian up to 7 percentage points on dative objects in German (compared with the PRED-M model). What we can see is that, in general, the improvements seem to be higher on the more infrequent arguments like dative objects and predicates than on the frequent arguments like subject or accusative object. It is not the case, however, that the performance of one of the infrequent functions suddenly surpasses the performance of a more frequent function. Those two effects are to be expected because the ILP parser is still a data-driven parser. The constraints support it by excluding morpho-syntactically incorrect analyses but they do not resolve ambiguous cases, which are still decided by the statistical model.

The main work done by the constraints is to establish interactions between parts of the parse graph that are not represented in the statistical model. Because the graph-based approach (in both parsers) factors the graph into first- (and some second-) order arcs, and because both decoders do not use second-order features with more than one label, a constraint like *label uniqueness* (Equation (10)), which is not even directly related to morphology, is impossible to learn for the statistical model. This is because it never sees two sister dependents and their labels together and thus does not know if it has already annotated the current function label. Applying the constraints during the search makes it impossible for the parser to produce an output that does not

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

obey label uniqueness even though the statistical model does not have access to this information.

It should be stressed that the ILP models in their statistical model still use the same predicted and fully disambiguated morphological information from the pipeline architecture as the Bohnet parser. As we saw in the first experiment, using no morphological information in the statistical model is very harmful to the performance on Czech and Hungarian, though not so much for German.

One advantage of the proposed architecture is the fact that the ILP parser is still mainly driven by the statistical model. Krivanek and Meurers (2011) compared a data-driven, transition-based dependency parser (Nivre et al. 2007b) and a constraint-based dependency parser (Foth and Menzel 2006) on learner and newspaper corpora and found that whereas the former is better on modifier functions (e.g., PP-attachment), the latter performs better on argument functions. Their explanation is that where the data-driven parser has access to lots of data and can pick up statistical effects in the data like semantic or selectional preferences, the constraint-based parser has access to deep lexical and grammatical information and is thus able to model argument structure in a better way. In the ILP parser, we can combine both strengths, letting the statistical model learn preferences but forcing it via constraints to obey hard grammatical information. The last row in Table 11 shows that compared to the Bohnet parser, the ILP models perform comparably well on non-argument functions (maybe with the exception of Hungarian, where the difference is a bit more distinct). At the same time, they perform clearly better on the argument functions due to the linguistic constraints.

Foth and Menzel (2006) (see also Khmylko, Foth, and Menzel 2009) are further relevant to this work in the sense that our architecture mirrors their approach. In their work, they use a highly sophisticated rule-based parser, which they equip with statistical components that model various subtasks like pos tagging, supertagging, or PP-attachment. They demonstrate that a rule-based parser can benefit from statistical models that model preferences rather than hard constraints. Our approach comes from the other side: We equip a statistical parser with hard rules that ensure the linguistic plausibility of the output. Both approaches prove that proper statistical models and linguistically motivated rules can work well together to produce syntactic structures of high quality.

One advantage of applying constraints over the argument structure is that we can give a guarantee that certain ill-formed trees will not be produced by the parser. For example, the constraints make sure that there will not be any parser output where there are two subjects annotated for the same verb. Although this does not mean that the subject will be the correct one, the formal requirement of not having two subjects is met, which we believe can be helpful for subsequent semantic analysis/interpretation or, for example, relation extraction. In the same sense, the constraints will also ensure that morphological agreement and case licensing is correct to the degree that the morphological analyzer was correct. This feature thus implements a tentative notion of grammaticality for the statistical model.

7. Conclusion

In this article, we investigated the performance of the state-of-the-art statistical dependency parser by Bohnet (2010) on three morphologically rich languages—Czech, German, and Hungarian. We concentrated on the core grammatical functions (subject,

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

object, etc.) that are marked by case morphology in each of the three languages. Our first experiment shows that apart from small frequency effects due to the statistical nature of the parser, learning the mapping between a case value and the grammatical functions signaled by it is not a problem for the parser. We also see, however, that the pipeline approach, where morphological information is fully disambiguated before being used by the parser as features in the statistical model, is susceptible to error propagation for languages that show syncretism in their morphological paradigms. Although we can show that parsing Hungarian, an agglutinating language without major syncretism in the case paradigm, is not affected by these problems, parsing the fusional languages Czech and German frequently suffers from propagated errors due to ambiguous case morphology. Furthermore, although the predicted morphological information does not help very much in German, it contributes very much when parsing Czech and Hungarian, even if it is not completely reliable.

Handling syncretism requires changes in the processing architecture and the representation of morphological information. We proposed an augmented pipeline where the parsing model is restricted by possibly underspecified, morpho-syntactic constraints exploiting grammatical knowledge about the morphological marking regimes and the inflectional paradigms. Although the statistical parsing model provides scores for local substructures during decoding, the symbolic constraints are applied globally to the entire output structure. A morpho-syntactic feature like case is interpreted as a filter on the parser output. By modeling phenomena like case-function mapping, agreement, and function uniqueness as constraints in an ILP decoder for dependency parsing, we showed in a second experiment that supporting a statistical model with these constraints helps avoiding parsing errors due to incorrect morphological preprocessing. The advantage of this approach is the combination of local statistical models and globally enforced hard grammatical knowledge. Whereas some key aspects of the grammatical structure are ensured by the linguistic knowledge (e.g., overtly marked case morphology) the underlying data-driven model can still exploit statistical effects to resolve the remaining ambiguity and model semantic preferences, which are difficult to model with hard rules.

Morphologically rich languages pose various challenges to the standard parsing approaches because of their different linguistic properties. As one of them, case systems are a key device in these languages to encode argument structure and reside at the brink between morphology and syntax. Paying attention to the role of case in statistical parsing results in more appropriate models. *Morphologically rich*, however, is a wide category and covers a wide range of languages. Taking the idea of linguistically informed restrictions over data-driven system components may lead to further improvements on other phenomena and for other languages.

Acknowledgments

The research reported in this article was supported by the German Research Foundation (DFG) in project D8 of SFB 732 *Incremental Specification in Context*. We would like to thank Richárd Farkas and Veronika Vincze at the University of Szeged for their help with the Hungarian corpus and language; Bernd Bohnet for the help with his parser; and Anders Björkelund, Anett Diesner, and Kyle Richardson for their comments on earlier drafts of this work.

References

Blake, Barry J. 2001. *Case*. Cambridge University Press, Cambridge, MA, 2nd edition.

Böhmová, Alena, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2000. The Prague Dependency Treebank: A three-level annotation scenario. In A. Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora.* Kluwer Academic Publishers, Amsterdam, chapter 1, pages 103–127.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Bohnet, Bernd. 2009. Efficient parsing of syntactic and semantic dependency structures. In *Proceedings of the 13th Conference on Computational Natural Language Learning: Shared Task*, volume 2007, pages 67–72, Boulder, CO.

Bohnet, Bernd. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing.

Bohnet, Bernd. 2011. Comparing advanced graph-based and transition-based dependency. In *Proceedings of the International Conference on Dependency Linguistics*, pages 282–289, Barcelona.

Boyd, Adriane, Markus Dickinson, and W. Detmar Meurers. 2008. On detecting errors in dependency treebanks. *Research on Language and Computation*, 6(2):113–137.

Brants, Sabine, Stefanie Dipper, Silvia Hansen-Shirra, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the 1st Workshop on Treebanks and Linguistic Theories*, 20–21 September 2002, Sozopol, Bulgaria, pages 24–41.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Blackwell Publishers, Oxford.

Buchholz, Sabine and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 149–164, New York, NY.

Carreras, Xavier. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 957–961, Prague.

Cohen, Shay B. and Noah A. Smith. 2007. Joint morphological and syntactic disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 208–217, Prague.

Collins, Michael, Jan Hajič, Lance Ramshaw, and Christoph Tillmann. 1999. A statistical parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, College Park, MD.

Crammer, Koby, Ofer Dekel, Shai Shalev-Shwartz, and Yoram Singer. 2003. Online passive-aggressive algorithms. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems*, volume 7, pages 1217–1224, Cambridge, MA.

Csendes, Dóra, János Csirik, and Tibor Gyimóthy. 2004. The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, pages 19–23, Geneva.

Eisenberg, Peter. 2006. *Grundriss der deutschen Grammatik: Der Satz*. J.B. Metzler, Stuttgart, 3rd edition.

Eisner, Jason. 1997. Bilexical grammars and a cubic-time probabilistic parser. In *Proceedings of the 5th International Conference on Parsing Technologies*, pages 54–65, Cambridge, MA.

Eryiğit, Gülşen, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

Foth, Kilian A. and Wolfgang Menzel. 2006. Hybrid parsing: Using probabilistic models as predictors for a symbolic parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 321–328, Sidney.

Gesmundo, Andrea, James Henderson, Paola Merlo, and Ivan Titov. 2009. A latent variable model of synchronous syntactic-semantic parsing for multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning: Shared Task*, pages 37–42, Boulder, CO.

Goldberg, Yoav and Michael Elhadad. 2010. Easy first dependency parsing of modern Hebrew. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 103–107, Los Angeles, CA.

Goldberg, Yoav and Reut Tsarfaty. 2008. A single generative model for joint morphological segmentation and syntactic parsing. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 371–379, Columbus, OH.

Hajič, Jan, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Stepánek, Pavel Stranák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural*

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Language Learning: Shared Task, pages 1–18, Boulder, CO.
Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank 2.0, Linguistic Data Consortium, Philadelphia, PA.
Hudson, Richard A. 1984. *Word Grammar*. Basil Blackwell, Oxford.
Janda, Laura A. and Charles E. Townsend. 2000. *Czech*. Lincom Europa, Munich.
Johansson, Richard and Pierre Nugues. 2008. Dependency-based syntactic-semantic analysis with PropBank and NomBank. In *Proceedings of the 12th Conference on Computational Natural Language Learning*, pages 183–187, Manchester.
Khmylko, Lidia, Kilian A. Foth, and Wolfgang Menzel. 2009. Co-parsing with competitive Models. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 99–107, Paris.
Klenner, Manfred. 2007. Shallow dependency labeling. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 201–204, Prague.
Krivanek, Julia and W. Detmar Meurers. 2011. Comparing rule-based and datadriven dependency parsing of learner language. In *Proceedings of the International Conference on Dependency Linguistics*, pages 310–318, Barcelona.
Kübler, Sandra. 2008. The PaGe 2008 shared task on parsing German. In *Proceedings of the Workshop on Parsing German*, pages 55–63, Morristown, NJ.
Lee, John, Jason Naradowsky, and David A. Smith. 2011. A discriminative model for joint morphological disambiguation and dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 885–894, Portland, OR.
Magnanti, Thomas and Laurence Wolsey. 1995. Optimal trees. *Handbooks in Operations Research and Management Science*, 7(April):503–615.
Martins, André F. T., Noah A. Smith, and Eric P. Xing. 2009. Concise integer linear programming formulations for dependency parsing. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 342–350, Suntec.
Marton, Yuval, Nizar Habash, and Owen Rambow. 2010. Improving Arabic dependency parsing with lexical and inflectional morphological features. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 13–21, Los Angeles, CA.
McDonald, Ryan and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 81–88, Trento.
McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Morristown, NJ.
Mel'čuk, Igor. 1988. *Dependency Syntax: Theory and Practice*. SUNY Series in Linguistics. State University Press of New York.
Nichols, Joanna. 1986. Head-marking and dependent-marking grammar. *Language*, 62(1):56–119.
Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007a. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 915–932, Prague.
Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007b. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.
Riedel, Sebastian and James Clarke. 2006. Incremental integer linear programming for non-projective dependency parsing. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 129–137, Sydney.
Schiehlen, Michael. 2004. Annotation strategies for probabilistic parsing in German. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 390–397, Geneva.
Schiller, Anne. 1994. Dmor - user's guide. Technical report, University of Stuttgart.
Seeker, Wolfgang and Jonas Kuhn. 2012. Making ellipses explicit in dependency conversion for a German treebank. In *Proceedings of the 8th International*

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Conference on Language Resources and Evaluation, pages 3132–3139, Istanbul.
Seeker, Wolfgang and Jonas Kuhn. 2011. On the role of explicit morphological feature representation in syntactic dependency parsing for German. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 58–62, Dublin.
Seeker, Wolfgang, Ines Rehbein, Jonas Kuhn, and Josef Van Genabith. 2010. Hard constraints for grammatical function labelling. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1087–1097, Uppsala.
Spoustová, Drahomíra "Johanka," Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 763–771, Athens.
Taskar, Ben, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. 2005. Learning structured prediction models: A large margin approach. In *Proceedings of the 22th Annual International Conference on Machine Learning*, pages 896–903, Bonn.
Trón, Viktor, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1670–1673, Genoa, Italy.
Tsarfaty, Reut, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL): What, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12, Los Angeles, CA.
Tsarfaty, Reut and Khalil Sima'an. 2008. Relational-realizational parsing. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 889–896, Manchester.
Tsarfaty, Reut and Khalil Sima'an. 2010. Modeling morphosyntactic agreement in constituency-based parsing of Modern Hebrew. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 40–48, Los Angeles, CA.
Versley, Yannick. 2005. Parser evaluation across text types. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories*, pages 209–220, Barcelona.
Versley, Yannick and Ines Rehbein. 2009. Scalable discriminative parsing for German. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 134–137, Paris.
Vincze, Veronika, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian Dependency Treebank. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, pages 1855–1862, Valletta.
Zsibrita, János, Veronika Vincze, and Richárd Farkas. 2010. Ismeretlen kifejezések és a szófaji egyértelmüsítés. In *VII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 275–283, Szeged.

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023

Downloaded from http://direct.mit.edu/coli/article-pdf/39/1/23/1799024/coli_a_00134.pdf by guest on 07 September 2023