Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither

Reut Tsarfaty
Uppsala Universitet

Djamé Seddah
Alpage (Inria/Univ. Paris-Sorbonne)

Yoav Goldberg
Ben Gurion University

Sandra Kübler
Indiana University

Marie Candito
Alpage (Inria/Univ. Paris 7)

Jennifer Foster
NCLT, Dublin City University

Yannick Versley
Universität Tübingen

Ines Rehbein
Universität Saarbrücken

Lamia Tounsi
NCLT, Dublin City University

Abstract

The term Morphologically Rich Languages (MRLs) refers to languages in which significant information concerning syntactic units and relations is expressed at word-level. There is ample evidence that the application of readily available statistical parsing models to such languages is susceptible to serious performance degradation. The first workshop on statistical parsing of MRLs hosts a variety of contributions which show that despite language-specific idiosyncrasies, the problems associated with parsing MRLs cut across languages and parsing frameworks. In this paper we review the current state-of-affairs with respect to parsing MRLs and point out central challenges. We synthesize the contributions of researchers working on parsing Arabic, Basque, French, German, Hebrew, Hindi and Korean to point out shared solutions across languages. The overarching analysis suggests itself as a source of directions for future investigations.

1 Introduction

The availability of large syntactically annotated corpora led to an explosion of interest in automatically inducing models for syntactic analysis and disambiguation called *statistical parsers*. The development of successful statistical parsing models for English focused on the Wall Street Journal Penn Treebank (PTB, (Marcus et al., 1993)) as the primary, and sometimes only, resource. Since the initial release of the Penn Treebank (PTB Marcus et al. (1993)), many different constituent-based parsing models have been developed in the context of parsing English (*e.g.* (Magerman, 1995; Collins, 1997; Charniak, 2000; Chiang, 2000; Bod, 2003; Charniak and Johnson, 2005; Petrov et al., 2006; Huang, 2008; Finkel et al., 2008; Carreras et al., 2008)). At their time, each of these models improved the state-of-the-art, bringing parsing performance on the standard test set of the Wall-Street-Journal to a performance ceiling of 92% F_1-score using the PARSEVAL evaluation metrics (Black et al., 1991). Some of these parsers have been adapted to other language/treebank pairs, but many of these adaptations have been shown to be considerably less successful.

Among the arguments that have been proposed to explain this performance gap are the impact of small data sets, differences in treebanks' annotation schemes, and inadequacy of the widely used PARSEVAL evaluation metrics. None of these aspects in isolation can account for the systematic performance deterioration, but observed from a wider, cross-linguistic perspective, a picture begins to emerge – that the morphologically rich nature of some of the languages makes them inherently more susceptible to such performance degradation. Linguistic factors associated with MRLs, such as a large inventory of word-forms, higher degrees of word order freedom, and the use of morphological information in indicating syntactic relations, makes them substantially harder to parse with models and techniques that have been developed with English data in mind.

Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 1–12,
Los Angeles, California, June 2010.

In addition to these technical and linguistic factors, the prominence of English parsing in the literature reduces the visibility of research aiming to solve problems particular to MRLs. The lack of streamlined communication among researchers working on different MRLs often leads to a *reinventing the wheel* syndrome. To circumvent this, the first workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010) offers a platform for this growing community to share their views of the different problems and oftentimes similar solutions.

We identify three main types of challenges, each of which raises many questions. Many of the questions are yet to be conclusively answered. The first type of challenges has to do with the architectural setup of parsing MRLs: *What is the nature of the input? Can words be represented abstractly to reflect shared morphological aspects? How can we cope with morphological segmentation errors propagated through the pipeline?* The second type concerns the representation of morphological information inside the articulated syntactic model: *Should morphological information be encoded at the level of PoS tags? On dependency relations? On top of non-terminals symbols? How should the integrated representations be learned and used?* A final genuine challenge has to do with sound estimation for lexical probabilities: *Given the finite, and often rather small, set of data, and the large number of morphological analyses licensed by rich inflectional systems, how can we analyze words unseen in the training data?*

Many of the challenges reported here are mostly irrelevant when parsing Section 23 of the PTB but they are of primordial importance in other tasks, including out-of-domain parsing, statistical machine translation, and parsing resource-poor languages. By synthesizing the contributions to the workshop and bringing it to the forefront, we hope to advance the state of the art of statistical parsing in general.

In this paper we therefore take the opportunity to analyze the knowledge that has been acquired in the different investigations for the purpose of identifying main bottlenecks and pointing out promising research directions. In section 2, we define MRLs and identify syntactic characteristics associated with them. We then discuss work on parsing MRLs in both the dependency-based and constituency-based setup. In section 3, we review the types of challenges associated with parsing MRLs across frameworks. In section 4, we focus on the contributions to the SPMRL workshop and identify recurring trends in the empirical results and conceptual solutions. In section 5, we analyze the emerging picture from a bird's eye view, and conclude that many challenges could be more faithfully addressed in the context of parsing morphologically ambiguous input.

2 Background

2.1 What are MRLs?

The term Morphologically Rich Languages (MRLs) is used in the CL/NLP literature to refer to languages in which substantial grammatical information, i.e., information concerning the arrangement of words into syntactic units or cues to syntactic relations, is expressed at word level.

The common linguistic and typological wisdom is that "morphology competes with syntax" (Bresnan, 2001). In effect, this means that *rich morphology* goes hand in hand with a host of *nonconfigurational* syntactic phenomena of the kind discussed by Hale (1983). Because information about the relations between syntactic elements is indicated in the form of words, these words can freely change their positions in the sentence. This is referred to as *free word order* (Mithun, 1992). Information about the grouping of elements together can further be expressed by reference to their morphological form. Such logical groupings of disparate elements are often called *discontinuous constituents*. In dependency structures, such discontinuities impose *nonprojectivity*. Finally, rich morphological information is found in abundance in conjunction with so-called *pro-drop* or *zero anaphora*. In such cases, rich morphological information in the head (or co-head) of the clause often makes it possible to omit an overt subject which would be semantically impoverished.

English, the most heavily studied language within the CL/NLP community, is not an MRL. Even though a handful of syntactic features (such as person and number) are reflected in the form of words, morphological information is often secondary to other syntactic factors, such as the position of words and their arrangement into phrases. German, an Indo-European language closely related to English, already exhibits some of the properties that make

parsing MRLs problematic. The Semitic languages Arabic and Hebrew show an even more extreme case in terms of the richness of their morphological forms and the flexibility in their syntactic ordering.

2.2 Parsing MRLs

Pushing the envelope of constituency parsing: The Head-Driven models of the type proposed by Collins (1997) have been ported to parsing many MRLs, often via the implementation of Bikel (2002). For Czech, the adaptation by Collins et al. (1999) culminated in an 80 F_1-score.

German has become almost an archetype of the problems caused by MRLs; even though German has a moderately rich morphology and a moderately free word order, parsing results are far from those for English (see (Kübler, 2008) and references therein). Dubey (2005) showed that, for German parsing, adding case and morphology information together with smoothed markovization and an adequate unknown-word model is more important than lexicalization (Dubey and Keller, 2003).

For Modern Hebrew, Tsarfaty and Sima'an (2007) show that a simple treebank PCFG augmented with parent annotation and morphological information as state-splits significantly outperforms Head-Driven markovized models of the kind made popular by Klein and Manning (2003). Results for parsing Modern Standard Arabic using Bikel's implementation on gold-standard tagging and segmentation have not improved substantially since the initial release of the treebank (Maamouri et al., 2004; Kulick et al., 2006; Maamouri et al., 2008).

For Italian, Corazza et al. (2004) used the Stanford parser and Bikel's parser emulation of Collins' model 2 (Collins, 1997) on the ISST treebank, and obtained significantly lower results compared to English. It is notable that these models were applied without adding morphological signatures, using gold lemmas instead. Corazza et al. (2004) further tried different refinements including parent annotation and horizontal markovization, but none of them obtained the desired improvement.

For French, Crabbé and Candito (2008) and Seddah et al. (2010) show that, given a corpus comparable in size and properties (*i.e.* the number of tokens and grammar size), the performance level, both for Charniak's parser (Charniak, 2000) and the Berkeley parser (Petrov et al., 2006) was higher for parsing the PTB than it was for French. The split-merge-smooth implementation of (Petrov et al., 2006) consistently outperform various lexicalized and unlexicalized models for French (Seddah et al., 2009) and for many other languages (Petrov and Klein, 2007). In this respect, (Petrov et al., 2006) is considered MRL-friendly, due to its language agnostic design.

The rise of dependency parsing: It is commonly assumed that dependency structures are better suited for representing the syntactic structures of free word order, morphologically rich, languages, because this representation format does not rely crucially on the position of words and the internal grouping of surface chunks (Mel'čuk, 1988). It is an entirely different question, however, whether dependency parsers are in fact better suited for parsing such languages.

The CoNLL shared tasks on multilingual dependency parsing in 2006 and 2007 (Buchholz and Marsi, 2006; Nivre et al., 2007a) demonstrated that dependency parsing for MRLs is quite challenging. While dependency parsers are adaptable to many languages, as reflected in the multiplicity of the languages covered,[1] the analysis by Nivre et al. (2007b) shows that the best result was obtained for English, followed by Catalan, and that the most difficult languages to parse were Arabic, Basque, and Greek. Nivre et al. (2007a) drew a somewhat typological conclusion, that languages with rich morphology and free word order are the hardest to parse. This was shown to be the case for both MaltParser (Nivre et al., 2007c) and MST (McDonald et al., 2005), two of the best performing parsers on the whole.

Annotation and evaluation matter: An emerging question is therefore whether models that have been so successful in parsing English are necessarily appropriate for parsing MRLs – but associated with this question are important questions concerning the annotation scheme of the related treebanks. Obviously, when annotating structures for languages with characteristics different than English one has to face different annotation decisions, and it comes as no surprise that the annotated structures for MRLs often differ from those employed in the PTB.

[1] The shared tasks involved 18 languages, including many MRLs such as Arabic, Basque, Czech, Hungarian, and Turkish.

For Spanish and French, it was shown by Cowan and Collins (2005) and in (Arun and Keller, 2005; Schluter and van Genabith, 2007), that restructuring the treebanks' native annotation scheme to match the PTB annotation style led to a significant gain in parsing performance of Head-Driven models of the kind proposed in (Collins, 1997). For German, a language with four different treebanks and two substantially different annotation schemes, it has been shown that a PCFG parser is sensitive to the kind of representation employed in the treebank.

Dubey and Keller (2003), for example, showed that a simple PCFG parser outperformed an emulation of Collins' model 1 on NEGRA. They showed that using sister-head dependencies instead of head-head dependencies improved parsing performance, and hypothesized that it is due to the flatness of phrasal annotation. Kübler et al. (2006) showed considerably lower PARSEVAL scores on NEGRA (Skut et al., 1998) relative to the more hierarchically structured TüBa-D/Z (Hinrichs et al., 2005), again, hypothesizing that this is due to annotation differences.

Related to such comparisons is the question of the relevance of the PARSEVAL metrics for evaluating parsing results across languages and treebanks. Rehbein and van Genabith (2007) showed that PARSEVAL measures are sensitive to annotation scheme particularities (*e.g.* the internal node ratio). It was further shown that different metrics (*i.e.* the Leaf-ancestor path (Sampson and Babarczy, 2003) and dependency based ones in (Lin, 1995)) can lead to different performance ranking. This was confirmed also for French by Seddah et al. (2009).

The questions of how to annotate treebanks for MRLs and how to evaluate the performance of the different parsers on these different treebanks is crucial. For the MRL parsing community to be able to assess the difficulty of improving parsing results for French, German, Arabic, Korean, Basque, Hindi or Hebrew, we ought to first address fundamental questions including: Is the treebank sufficiently large to allow for proper grammar induction? Does the annotation scheme fit the language characteristics? Does the use of PTB annotation variants for other languages influence parsing results? Does the space-delimited tokenization allow for phrase boundary detection? Do the results for a specific approach generalize to more than one language?

3 Primary Research Questions

It is firmly established in theoretical linguistics that morphology and syntax closely interact through patterns of case marking, agreement, clitics and various types of compounds. Because of such close interactions, we expect morphological cues to help parsing performance. But in practice, when trying to incorporate morphological information into parsing models, three types of challenges present themselves:

Architecture and Setup: When attempting to parse complex word-forms that encapsulate both lexical and functional information, important architectural questions emerge, namely, what is the nature of the input that is given to the parsing system? Does the system attempt to parse sequences of words or does it aim to assign structures to sequences of morphological segments? If the former is the case, how can we represent words abstractly so as to reflect shared morphological aspects between them? If the latter is the case, how can we arrive at a good enough morphological segmentation for the purpose of statistical parsing, given raw input texts?

When working with morphologically rich languages such as Hebrew or Arabic, affixes may have syntactically independent functions. Many parsing models assume segmentation of the syntactically independent parts, such as prepositions or pronominal clitics, prior to parsing. But morphological segmentation requires disambiguation which is non-trivial, due to case syncretism and high morphological ambiguity exhibited by rich inflectional systems. The question is then when should we disambiguate the morphological analyses of input forms? Should we do that prior to parsing or perhaps jointly with it?[2]

Representation and Modeling: Assuming that the input to our system reflects morphological information, one way or another, which *types* of morpho-

[2]Most studies on parsing MRLs nowadays assume the gold standard segmentation and disambiguated morphological information as input. This is the case, for instance, for the Arabic parsing at CoNLL 2007 (Nivre et al., 2007a). This practice deludes the community as to the validity of the parsing results reported for MRLs in shared tasks. Goldberg et al. (2009), for instance, show a gap of up to 6pt F_1-score between performance on gold standard segmentation vs. raw text. One way to overcome this is to devise joint morphological and syntactic disambiguation frameworks (cf. (Goldberg and Tsarfaty, 2008)).

logical information should we include in the parsing model? Inflectional and/or derivational? Case information and/or agreement features? How can valency requirements reflected in derivational morphology affect the overall syntactic structure? In tandem with the decision concerning the morphological information to include, we face genuine challenges concerning how to represent such information in the syntactic model, be it constituency-based or dependency-based. Should we encode morphological information at the level of PoS tags and/or on top of syntactic elements? Should we decorate non-terminals nodes and/or dependency arcs or both?

Incorporating morphology in the statistical model is often even more challenging than the sum of these bare decisions, because of the nonconfigurational structures (free word order, discontinuous constituents) for rich markings are crucial (Hale, 1983). The parsing models designed for English often focus on learning rigid word order, and they do not take morphological information into account (cf. developing parsers for German (Dubey and Keller, 2003; Kübler et al., 2006)). The more complex question is therefore: what type of parsing model should we use for parsing MRLs? shall we use a general purpose implementation and attempt to amend it? how? or perhaps we should devise a new model from first principles, to address nonconfigurational phenomena effectively? using what form of representation? is it possible to find a single model that can effectively cope with different kinds of languages?

Estimation and Smoothing: Compared to English, MRLs tend to have a greater number of word forms and higher out-of-vocabulary (OOV) rates, due to the many feature combinations licensed by the inflectional system. A typical problem associated with parsing MRLs is substantial lexical data sparseness due to high morphological variation in surface forms. The question is therefore, given our finite, and often fairly small, annotated sets of data, how can we guess the morphological analyses, including the PoS tag assignment and various features, of an OOV word? How can we learn the probabilities of such assignments? In a more general setup, this problem is akin to handling out-of-vocabulary or rare words for robust statistical parsing, and techniques for domain adaptation via lexicon enhancement (also explored for English and other morphologically impoverished languages).

So, in fact, incorporating morphological information inside the syntactic model for the purpose of statistical parsing is anything but trivial. In the next section we review the various approaches taken in the individual contributions of the SPMRL workshop for addressing such challenges.

	Constituency-Based	Dependency-Based
Arabic	(Attia et al., 2010)	(Marton et al., 2010)†
Basque	-	(Bengoetxea and Gojenola, 2010)
English	(Attia et al., 2010)	-
French	(Attia et al., 2010)	
	(Seddah et al., 2010)	
	(Candito and Seddah, 2010)†	-
German	(Maier, 2010)	-
Hebrew	(Tsarfaty and Sima'an, 2010)	(Goldberg and Elhadad, 2010)†
Hindi	-	(Ambati et al., 2010a)†
		(Ambati et al., 2010b)
Korean	(Chung et al., 2010)	-

Table 1: An overview of SPMRL contributions. († report results also for non-gold standard input)

4 Parsing MRLs: Recurring Trends

The first workshop on parsing MRLs features 11 contributions for a variety of languages with a range of different parsing frameworks. Table 1 lists the individual contributions within a cross-language cross-framework grid. In this section, we focus on trends that occur among the different contributions. This may be a biased view since some of the problems that exist for parsing MRLs may have not been at all present, but it is a synopsis of where we stand with respect to problems that are being addressed.

4.1 Architecture and Setup: Gold vs. Predicted Morphological Information

While morphological information can be very informative for syntactic analysis, morphological analysis of surface forms is ambiguous in many ways. In German, for instance, case syncretism (*i.e.* a single surface form corresponding to different cases) is pervasive, and in Hebrew and Arabic, the lack of vocalization patterns in written texts leads to multiple morphological analyses for each space-delimited token. In real world situations, gold morphological information is not available prior to parsing. Can parsing systems make effective use of morphology even when gold morphological information is absent?

Several papers address this challenge by presenting results for both the gold and the automatically predicted PoS and morphological information (Ambati et al., 2010a; Marton et al., 2010; Goldberg and Elhadad, 2010; Seddah et al., 2010). Not very surprisingly, all evaluated systems show a drop in parsing accuracy in the non-gold settings.

An interesting trend is that in many cases, using noisy morphological information is worse than not using any at all. For Arabic Dependency parsing, using predicted CASE causes a substantial drop in accuracy while it greatly improves performance in the gold setting (Marton et al., 2010). For Hindi Dependency Parsing, using chunk-internal cues (*i.e.* marking non-recursive phrases) is beneficial when gold chunk-boundaries are available, but suboptimal when they are automatically predicted (Ambati et al., 2010a). For Hebrew Dependency Parsing with the MST parser, using gold morphological features shows no benefit over not using them, while using automatically predicted morphological features causes a big drop in accuracy compared to not using them (Goldberg and Elhadad, 2010). For French Constituency Parsing, Seddah et al. (2010) and Candito and Seddah (2010) show that while gold information for the part-of-speech and lemma of each word form results in a significant improvement, the gain is low when switching to predicted information. Reassuringly, Ambati et al. (2010a), Marton et al. (2010), and Goldberg and Elhadad (2010) demonstrate that some morphological information can indeed be beneficial for parsing even in the automatic setting. Ensuring that this is indeed so, appears to be in turn linked to the question of how morphology is represented and incorporated in the parsing model.

The same effect in a different guise appears in the contribution of Chung et al. (2010) concerning parsing Korean. Chung et al. (2010) show a significant improvement in parsing accuracy when including traces of null anaphors (a.k.a. *pro-drop*) in the input to the parser. Just like overt morphology, traces and null elements encapsulate functional information about relational entities in the sentence (the subject, the object, etc.), and including them at the input level provides helpful disambiguating cues for the overall structure that represents such relations. However, assuming that such traces are given prior to parsing is, for all practical purposes, infeasible. This leads to an interesting question: will identifying such functional elements (marked as traces, overt morphology, etc) *during* parsing, while complicating that task itself, be on the whole justified?

Closely linked to the inclusion of morphological information in the input is the choice of PoS tag set to use. The generally accepted view is that fine-grained PoS tags are morphologically more informative but may be harder to statistically learn and parse with, in particular in the non-gold scenario. Marton et al. (2010) demonstrate that a fine-grained tag set provides the best results for Arabic dependency parsing when gold tags are known, while a much smaller tag set is preferred in the automatic setting.

4.2 Representation and Modeling: Incorporating Morphological Information

Many of the studies presented here explore the use of feature representation of morphological information for the purpose of syntactic parsing (Ambati et al., 2010a; Ambati et al., 2010b; Bengoetxea and Gojenola, 2010; Goldberg and Elhadad, 2010; Marton et al., 2010; Tsarfaty and Sima'an, 2010). Clear trends among the contributions emerge concerning the *kind* of morphological information that helps statistical parsing. Morphological CASE is shown to be beneficial across the board. It is shown to help for parsing Basque, Hebrew, Hindi and to some extent Arabic.[3] Morphological DEFINITENESS and STATE are beneficial for Hebrew and Arabic when explicitly represented in the model. STATE, ASPECT and MOOD are beneficial for Hindi, but only marginally beneficial for Arabic. CASE and SUBORDINATION-TYPE are the most beneficial features for Basque transition-based dependency parsing.

A closer view into the results mentioned in the previous paragraph suggests that, beyond the kind of information that is being used, the way in which morphological information is represented and used by the model has substantial ramification as to whether or not it leads to performance improvements. The so-called "agreement features" GENDER, NUMBER, PERSON, provide for an interesting case study in this respect. When included directly as

[3]For Arabic, CASE is useful when gold morphology information is available, but substantially hurt results when it is not.

machine learning features, agreement features benefit dependency parsing for Arabic (Marton et al., 2010), but not Hindi (dependency) (Ambati et al., 2010a; Ambati et al., 2010b) or Hebrew (Goldberg and Elhadad, 2010). When represented as simple splits of non-terminal symbols, agreement information does not help constituency-based parsing performance for Hebrew (Tsarfaty and Sima'an, 2010). However, when agreement patterns are directly represented on dependency arcs, they contribute an improvement for Hebrew dependency parsing (Goldberg and Elhadad, 2010). When agreement is encoded at the realization level inside a Relational-Realizational model (Tsarfaty and Sima'an, 2008), agreement features improve the state-of-the-art for Hebrew parsing (Tsarfaty and Sima'an, 2010).

One of the advantages of the latter study is that morphological information which is expressed at the level of words gets interpreted elsewhere, on functional elements higher up the constituency tree. In dependency parsing, similar cases may arise, that is, morphological information might not be as useful on the form on which it is expressed, but would be more useful at a different position where it could influence the correct attachment of the main verb to other elements. Interesting patterns of that sort occur in Basque, where the SUBORDINATIONTYPE morpheme attaches to the auxiliary verb, though it mainly influences attachments to the main verb.

Bengoetxea and Gojenola (2010) attempted two different ways to address this, one using a transformation segmenting the relevant morpheme and attaching it to the main verb instead, and another by propagating the morpheme along arcs, through a "stacking" process, to where it is relevant. Both ways led to performance improvements. The idea of a segmentation transformation imposes non-trivial pre-processing, but it may be that automatically learning the propagation of morphological features is a promising direction for future investigation.

Another, albeit indirect, way to include morphological information in the parsing model is using so-called latent information or some mechanism of clustering. The general idea is the following: when morphological information is added to standard terminal or non-terminal symbols, it imposes restrictions on the distribution of these no-longer-equivalent elements. Learning latent information does not represent morphological information directly, but presumably, the distributional restrictions can be automatically learned along with the splits of labels symbols in models such as (Petrov et al., 2006). For Korean (Chung et al., 2010), latent information contributes significant improvements. One can further do the opposite, namely, merging terminals symbols for the purpose of obtaining an abstraction over morphological features. When such clustering uses a morphological signature of some sort, it is shown to significantly improve constituency-based parsing for French (Candito and Seddah, 2010).

4.3 Representation and Modeling: Free Word Order and Flexible Constituency Structure

Off-the-shelf parsing tools are found in abundance for English. One problematic aspect of using them to parse MRLs lies in the fact that these tools focus on the statistical modeling of *configurational* information. These models often condition on the position of words relative to one another (*e.g.* in transition-based dependency parsing) or on the distance between words inside constituents (*e.g.* in Head-Driven parsing). Many of the contributions to the workshop show that working around existing implementations may be insufficient, and we may have to come up with more radical solutions.

Several studies present results that support the conjecture that when free word-order is explicitly taken into account, morphological information is more likely to contribute to parsing accuracy. The Relational-Realizational model used in (Tsarfaty and Sima'an, 2010) allows for reordering of constituents at a configuration layer, which is independent of the realization patterns learned from the data (*vis-à-vis* case marking and agreement). The easy-first algorithm of (Goldberg and Elhadad, 2010) which allows for significant flexibility in the order of attachment, allows the model to benefit from agreement patterns over dependency arcs that are easier to detect and attach first. The use of larger subtrees in (Chung et al., 2010) for parsing Korean, within a Bayesian framework, allows the model to learn distributions that take more elements into account, and thus learn the different distributions associated with morphologically marked elements in constituency structures, to improve performance.

In addition to free word order, MRLs show higher degree of freedom in extraposition. Both of these phenomena can result in discontinuous structures. In constituency-based treebanks, this is either annotated as additional information which has to be recovered somehow (traces in the case of the PTB, complex edge labels in the German TüBa-D/Z), or as discontinuous phrase structures, which cannot be handled with current PCFG models. Maier (2010) suggests the use of Linear Context-Free Rewriting Systems (LCFRSs) in order to make discontinuous structure transparent to the parsing process and yet preserve familiar notions from constituency.

Dependency representation uses non-projective dependencies to reflect discontinuities, which is problematic to parse with models that assume projectivity. Different ways have been proposed to deal with non-projectivity (Nivre and Nilsson, 2005; McDonald et al., 2005; McDonald and Pereira, 2006; Nivre, 2009). Bengoetxea and Gojenola (2010) discuss non-projective dependencies in Basque and show that the pseudo-projective transformation of (Nivre and Nilsson, 2005) improves accuracy for dependency parsing of Basque. Moreover, they show that in combination with other transformations, it improves the utility of these other ones, too.

4.4 Estimation and Smoothing: Coping with Lexical Sparsity

Morphological word form variation augments the vocabulary size and thus worsens the problem of lexical data sparseness. Words occurring with medium-frequency receive less reliable estimates, and the number of rare/unknown words is increased. One way to cope with the one of both aspects of this problem is through *clustering*, that is, providing an abstract representation over word forms that reflects their shared morphological and morphosyntactic aspects. This was done, for instance, in previous work on parsing German. Versley and Rehbein (2009) cluster words according to linear context features. These clusters include valency information added to verbs and morphological features such as case and number added to pre-terminal nodes. The clusters are then integrated as features in a discriminative parsing model to cope with unknown words. Their discriminative model thus obtains state-of-the-art results on parsing German.

Several contribution address similar challenges. For constituency-based generative parsers, the simple technique of replacing word forms with more abstract symbols is investigated by (Seddah et al., 2010; Candito and Seddah, 2010). For French, replacing each word form by its predicted part-of-speech and lemma pair results in a slight performance improvement (Seddah et al., 2010). When words are clustered, even according to a very local linear-context similarity measure, measured over a large raw corpus, and when word clusters are used in place of word forms, the gain in performance is even higher (Candito and Seddah, 2010). In both cases, the technique provides more reliable estimates for in-vocabulary words, since a given lemma or cluster appear more frequently. It also increases the known vocabulary. For instance, if a plural form is unseen in the training set but the corresponding singular form is known, then in a setting of using lemmas in terminal symbols, both forms are known.

For dependency parsing, Marton et al. (2010) investigates the use of morphological features that involve some semantic abstraction over Arabic forms. The use of undiacritized lemmas is shown to improve performance. Attia et al. (2010) specifically address the handling of unknown words in the latent-variable parsing model. Here again, the technique that is investigated is to project unknown words to more general symbols using morphological clues. A study on three languages, English, French and Arabic, shows that this method helps in all cases, but that the greatest improvement is obtained for Arabic, which has the richest morphology among three.

5 Where we're at

It is clear from the present overview that we are yet to obtain a complete understanding concerning which models effectively parse MRLs, how to annotate treebanks for MRLs and, importantly, how to evaluate parsing performance across types of languages and treebanks. These foundational issues are crucial for deriving more conclusive recommendations as to the kind of models and morphological features that can lead to advancing the state-of-the-art for parsing MRLs. One way to target such an understanding would be to encourage the investigation of particular tasks, individually or in the context

of shared tasks, that are tailored to treat those problematic aspects of MRLs that we surveyed here.

So far, constituency-based parsers have been assessed based on their performance on the PTB (and to some extent, across German treebanks (Kübler, 2008)) whereas comparison across languages was rendered opaque due to data set differences and representation idiosyncrasies. It would be interesting to investigate such a cross-linguistic comparison of parsers in the context of a shared task on constituency-based statistical parsing, in additional to dependency-based ones as reported in (Nivre et al., 2007a). Standardizing data sets for a large number of languages with different characteristics, would require us, as a community, to aim for constituency-representation guidelines that can represent the shared aspects of structures in different languages, while at the same time allowing differences between them to be reflected in the model.

Furthermore, it would be a good idea to introduce parsing tasks, for either constituent-based or dependency-based setups, which consider raw text as input, rather than morphologically segmented and analyzed text. Addressing the parsing problem while facing the morphological disambiguation challenge in its full-blown complexity would be illuminating and educating for at least two reasons: firstly, it would give us a better idea of what is the state-of-the-art for parsing MRLs in realistic scenarios. Secondly, it might lead to profound insights about the potentially successful ways to use morphology inside a parser, which may differ from the insights concerning the use of morphology in the less realistic parsing scenarios, where gold morphological information is given.

Finally, to be able to perceive where we stand with respect to parsing MRLs and how models fare against one another across languages, it would be crucial to arrive at evaluation metrics that capture information that is shared among the different representations, for instance, functional information concerning predicate-argument relations. Using the different kinds of measures in the context of cross-framework tasks will help us understand the utility of the different evaluation metrics that have been proposed and to arrive at a clearer picture of what it is that we wish to compare, and how we can faithfully do so across models, languages and treebanks.

6 Conclusion

This paper presents the synthesis of 11 contributions to the first workshop on statistical parsing for morphologically rich languages. We have shown that architectural, representational, and estimation issues associated with parsing MRLs are found to be challenging across languages and parsing frameworks. The use of morphological information in the non gold-tagged input scenario is found to cause substantial differences in parsing performance, and in the kind of morphological features that lead to performance improvements.

Whether or not morphological features help parsing also depends on the kind of model in which they are embedded, and the different ways they are treated within. Furthermore, sound statistical estimation methods for morphologically rich, complex lexica, turn out to be crucial for obtaining good parsing accuracy when using general-purpose models and algorithms. In the future we hope to gain better understanding of the common pitfalls in, and novel solutions for, parsing morphologically ambiguous input, and to arrive at principled guidelines for selecting the model and features to include when parsing different kinds of languages. Such insights may be gained, among other things, in the context of more morphologically-aware shared parsing tasks.

Acknowledgements

The program committee would like to thank NAACL for hosting the workshop and SIGPARSE for their sponsorship. We further thank INRIA Alpage team for their generous sponsorship. We are finally grateful to our reviewers and authors for their dedicated work and individual contributions.

References

Bharat Ram Ambati, Samar Husain, Sambhav Jain, Dipti Misra Sharma, and Rajeev Sangal. 2010a. Two methods to incorporate local morphosyntactic features in Hindi dependency parsing. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Bharat Ram Ambati, Samar Husain, Joakim Nivre, and Rajeev Sangal. 2010b. On the role of morphosyntactic features in Hindi dependency parsing. In *Proceedings*

of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010), Los Angeles, CA.

Abhishek Arun and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 306–313, Ann Arbor, MI.

Mohammed Attia, Jennifer Foster, Deirdre Hogan, Joseph Le Roux, Lamia Tounsi, and Josef van Genabith. 2010. Handling unknown words in statistical latent-variable parsing models for Arabic, English and French. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Kepa Bengoetxea and Koldo Gojenola. 2010. Application of different techniques to dependency parsing of Basque. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Daniel M. Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 178–182. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 306–311, San Mateo (CA). Morgan Kaufman.

Rens Bod. 2003. An efficient implementation of a new DOP model. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 19–26, Budapest, Hungary.

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

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

Marie Candito and Djamé Seddah. 2010. Parsing word clusters. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Xavier Carreras, Michael Collins, and Terry Koo. 2008. TAG, dynamic programming, and the perceptron for efficient, feature-rich parsing. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL)*, pages 9–16, Manchester, UK.

Eugene Charniak and Mark Johnson. 2005. Course-to-fine n-best-parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 173–180, Barcelona, Spain, June.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*, Seattle.

David Chiang. 2000. Statistical parsing with an automatically-extracted Tree Adjoining Grammar. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 456–463, Hong Kong. Association for Computational Linguistics Morristown, NJ, USA.

Tagyoung Chung, Matt Post, and Daniel Gildea. 2010. Factors affecting the accuracy of Korean parsing. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

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

Michael Collins. 1997. Three Generative, Lexicalized Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 16–23, Madrid, Spain.

Anna Corazza, Alberto Lavelli, Giogio Satta, and Roberto Zanoli. 2004. Analyzing an Italian treebank with state-of-the-art statistical parsers. In *Proceedings of the Third Third Workshop on Treebanks and Linguistic Theories (TLT 2004)*, Tübingen, Germany.

Brooke Cowan and Michael Collins. 2005. Morphology and reranking for the statistical parsing of Spanish. In *in Proceedins of EMNLP*.

Benoit Crabbé and Marie Candito. 2008. Expériences d'analyse syntaxique statistique du français. In *Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08)*, pages 45–54, Avignon, France.

Amit Dubey and Frank Keller. 2003. Probabilistic parsing for German using sister-head dependencies. In *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 96–103, Ann Arbor, MI.

Amit Dubey. 2005. What to do when lexicalization fails: parsing German with suffix analysis and smoothing. In *43rd Annual Meeting of the Association for Computational Linguistics*.

Jenny Rose Finkel, Alex Kleeman, and Christopher D. Manning. 2008. Efficient, feature-based, conditional random field parsing. In *Proceedings of ACL*.

Yoav Goldberg and Michael Elhadad. 2010. Easy-first dependency parsing of Modern Hebrew. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Yoav Goldberg and Reut Tsarfaty. 2008. A single framework for joint morphological segmentation and syntactic parsing. In *Proceedings of the 46nd Annual Meeting of the Association for Computational Linguistics*.

Yoav Goldberg, Reut Tsarfaty, Meni Adler, and Michael Elhadad. 2009. Enhancing unlexicalized parsing performance using a wide coverage lexicon, fuzzy tag-set mapping, and em-hmm-based lexical probabilities. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 327–335.

Kenneth L. Hale. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory*, 1(1).

Erhard W. Hinrichs, Sandra Kübler, and Karin Naumann. 2005. A unified representation for morphological, syntactic, semantic, and referential annotations. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 13–20, Ann Arbor, MI.

Liang Huang. 2008. Forest reranking: Discriminative parsing with non-local features. In *Proceedings of ACL*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430.

Sandra Kübler, Erhard W. Hinrichs, and Wolfgang Maier. 2006. Is it really that difficult to parse German? In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 111–119, Sydney, Australia, July. Association for Computational Linguistics.

Sandra Kübler. 2008. The PaGe 2008 shared task on parsing German. In *Proceedings of the Workshop on Parsing German*, pages 55–63. Association for Computational Linguistics.

Seth Kulick, Ryan Gabbard, and Mitchell Marcus. 2006. Parsing the Arabic treebank: Analysis and improvements. In *Proceedings of TLT*.

Dekang Lin. 1995. A dependency-based method for evaluating broad-coverage parsers. In *International Joint Conference on Artificial Intelligence*, pages 1420–1425, Montreal.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic treebank: Building a large-scale annotated Arabic corpus. In *Proceedings of NEMLAR International Conference on Arabic Language Resources and Tools*.

Mohamed Maamouri, Ann Bies, and Seth Kulick. 2008. Enhanced annotation and parsing of the Arabic treebank. In *Proceedings of INFOS*.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pages 276–283, Cambridge, MA.

Wolfgang Maier. 2010. Direct parsing of discontinuous constituents in german. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Yuval Marton, Nizar Habash, and Owen Rambow. 2010. Improving Arabic dependency parsing with lexical and inflectional morphological features. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Ryan T. McDonald and Fernando C. N. Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proc. of EACL'06*.

Ryan T. McDonald, Koby Crammer, and Fernando C. N. Pereira. 2005. Online large-margin training of dependency parsers. In *Proc. of ACL'05*, Ann Arbor, USA.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.

Marianne Mithun. 1992. Is basic word order universal? In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*. John Benjamins, Amsterdam.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI.

Joakim Nivre, 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 CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic, June. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007b. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007c. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Joakim Nivre. 2009. Non-projective dependency parsing in expected linear time. 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*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, July. Association for Computational Linguistics.

Ines Rehbein and Josef van Genabith. 2007. Treebank annotation schemes and parser evaluation for German. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic.

Geoffrey Sampson and Anna Babarczy. 2003. A test of the leaf-ancestor metric for parse accuracy. *Natural Language Engineering*, 9(04):365–380.

Natalie Schluter and Josef van Genabith. 2007. Preparing, restructuring, and augmenting a French Treebank: Lexicalised parsers or coherent treebanks? In *Proc. of PACLING 07*, Melbourne, Australia.

Djamé Seddah, Marie Candito, and Benoit Crabbé. 2009. Cross parser evaluation and tagset variation: A French Treebank study. In *Proceedings of the 11th Internation Conference on Parsing Technologies (IWPT'09)*, pages 150–161, Paris, France, October. Association for Computational Linguistics.

Djamé Seddah, Grzegorz Chrupała, Ozlem Cetinoglu, Josef van Genabith, and Marie Candito. 2010. Lemmatization and statistical lexicalized parsing of morphologically-rich languages. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper texts. In *ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany.

Reut Tsarfaty and Khalil Sima'an. 2007. Three-dimensional parametrization for parsing morphologically rich languages. In *Proceedings of the 10th International Conference on Parsing Technologies (IWPT)*, pages 156–167.

Reut Tsarfaty and Khalil Sima'an. 2008. Relational-Realizational parsing. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 889–896.

Reut Tsarfaty and Khalil Sima'an. 2010. Modeling morphosyntactic agreement in constituency-based parsing of Modern Hebrew. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Yannick Versley and Ines Rehbein. 2009. Scalable discriminative parsing for german. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 134–137, Paris, France, October. Association for Computational Linguistics.