Representation of Morphosyntactic Units and Coordination Structures in the Turkish Dependency Treebank

Umut Sulubacak **Gülşen Eryiğit**
Department of Computer Engineering
Istanbul Technical University
Istanbul, 34469, Turkey
{sulubacak, gulsen.cebiroglu}@itu.edu.tr

Abstract

This paper presents our preliminary conclusions as part of an ongoing effort to construct a new dependency representation framework for Turkish. We aim for this new framework to accommodate the highly agglutinative morphology of Turkish as well as to allow the annotation of unedited web data, and shape our decisions around these considerations. In this paper, we firstly describe a novel syntactic representation for morphosyntactic sub-word units (namely inflectional groups (IGs) in Turkish) which allows inter-IG relations to be discerned with perfect accuracy without having to hide lexical information. Secondly, we investigate alternative annotation schemes for coordination structures and present a better scheme (nearly 11% increase in recall scores) than the one in Turkish Treebank (Oflazer et al., 2003) for both parsing accuracies and compatibility for colloquial language.

1 Introduction

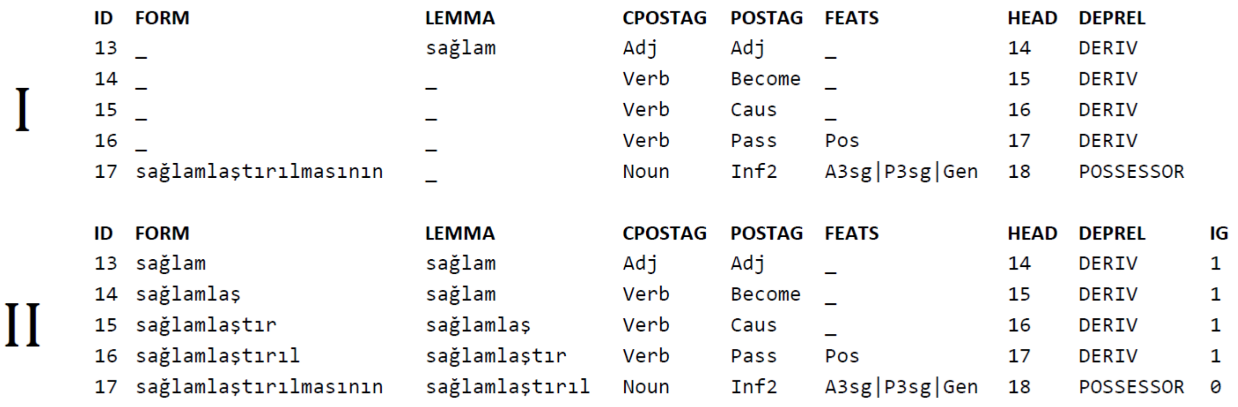
In recent years, dependency parsing has globally seen great deal of attention, and has constituted the underlying framework for the syntactic parsing of many multilingual studies. Even though constituency parsing and grammars are still the preferred formalism for some well-researched languages, others may have certain traits that put constituency parsing in an unfavorable position against dependency parsing, such as flexible constituent ordering, which is typical of several prominent languages including Turkish. Although Turkish is decidedly more workable over the dependency formalism, it has invariably fallen short of usual parsing accuracies compared to other languages, as seen clearly in some recent works such as (McDonald and Nivre, 2011).

There are more parameters to parsing than the formalism alone, among which the correctness of the corpora used in learning procedures and the annotation schemes of syntactic relations are held in consideration as part of this work. Between the two, the emphasis is on the annotation scheme, which is proven to significantly affect the parsing performance (Bosco et al., 2010; Boyd and Meurers, 2008). Our motivation for this research is that these factors must also contribute to some extent to the performance deficiency in parsing Turkish, besides the inherent difficulty of parsing the language. Our aim is to investigate these points and suggest improvements where applicable.

2 Parsing Framework and Data Set

As our parsing framework, we use MaltParser (Nivre et al., 2007) which is a data-driven dependency parser with an underlying SVM learner based on LIBSVM (Chang and Lin, 2001). MaltParser is widely used and has shown high performances across various languages (Nivre et al., 2006a). We run MaltParser with Nivre's Arc-Standard parsing algorithm (Nivre, 2003) and use the same optimized parameters as in (Eryiğit et al., 2008). We also use the learning features from the last cited work as our baseline feature set and an updated version from (Eryiğit et al., 2011) of the same data set (Oflazer, 2003). The only difference from the configuration of (Eryiğit et al., 2011) is that our baseline parser does not exclude non-projective sentences from the corpus for training, which explains the baseline ac-

Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages, pages 129–134,
Seattle, Washington, USA, 18 October 2013. ©2013 Association for Computational Linguistics



I

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
13	_	sağlam	Adj	Adj	_	14	DERIV
14	_	_	Verb	Become	_	15	DERIV
15	_	_	Verb	Caus	_	16	DERIV
16	_	_	Verb	Pass	Pos	17	DERIV
17	sağlamlaştırılmasının	_	Noun	Inf2	A3sg|P3sg|Gen	18	POSSESSOR

II

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL	IG
13	sağlam	sağlam	Adj	Adj	_	14	DERIV	1
14	sağlamlaş	sağlam	Verb	Become	_	15	DERIV	1
15	sağlamlaştır	sağlamlaş	Verb	Caus	_	16	DERIV	1
16	sağlamlaştırıl	sağlamlaştır	Verb	Pass	Pos	17	DERIV	1
17	sağlamlaştırılmasının	sağlamlaştırıl	Noun	Inf2	A3sg|P3sg|Gen	18	POSSESSOR	0

Figure 1: The original and the novel IG representations for the word *sağlamlaştırılmasının*, which respectively comes to mean *strong*, *to become strong*, *to strengthen*, *to be strengthened* and *of the strengthening of* after each derivation. The new morphological tags introduced after each derivation pertain to the relevant IG, and common morphological features for the IGs of a single word such as the agreement are given under the final IG. Model I is the original representation, while Model II is the new representation we propose.

curacy differences (e.g. 67.4% against our 65.0% in labelled attachment score).

3 Proposed Annotation Schemes

3.1 IGs

Within the context of data-driven parsing, the most apparent problem of languages with productive derivational morphology is that words can potentially yield a very large morphological tag set, which causes severe sparsity in the morphological features of words. To alleviate this problem, words are split into morphosyntactic parts called inflectional groups (IGs), taking intermediate derivational affixes as boundaries. It is a known fact that analyzing sentences as being composed of IGs rather than surface word forms yields better results in major NLP problems such as morphological disambiguation (Hakkani-Tür et al., 2002) and syntactic parsing (Eryiğit et al., 2008).

Within the domain of dependency parsing, IGs as syntactic tokens are not as free as independent words, since the IGs of each word must be connected to each other with an exclusive dependency relation named DERIV. However, other tokens are free to be connected to an arbitrary IG of a word, with the added benefit of more compact morphological feature sets to help make the distinction.

Other languages with productive derivation, such as Uralic or Ugric languages, or those orthographically differing from the well-studied European languages, such as Semitic languages, can also benefit from using non-word-based morphosyntactic parsing tokens, as evidenced for instance by the recent considerations of splitting up tokens based on morphemes for Hebrew (Tsarfaty and Goldberg, 2008).

3.1.1 Current IG Representation

Since MaltParser accepts input in the standard data format of the CoNLL-X Shared Task (Buchholz and Marsi, 2006), the ways in which IGs can be represented for the parser are limited. The standard method for annotating IGs using the CoNLL-X data fields, as described in (Eryiğit et al., 2008), involves marking up the FORM and LEMMA fields with underscores rather than with lexical data as shown in Figure 1. At first, this method is convenient, as current feature vectors readily take lexical information into account, and as such, a linear transition-based parser would easily learn to connect adjacent words as IGs of the same word as long as the head word has an underscore for a stem. However, an obvious drawback is that the actual lexical information gets lost in favor of marking IGs, preventing the potential usage of that information in deciding on inter-word dependencies.

3.1.2 Proposed IG Representation

As an improvement over the original IG representation described in Section 3.1.1, we propose a slightly different annotation scheme which does not lock out the lexical data columns, by making use of a new column named IG. This new column takes a boolean value that is true for non-final IGs of multi-IG words much like the original FORM column, effectively marking the dependents that must be connected to the next token in line with the dependency relation DERIV. Once this representation gets integrated, lexical information may be assigned to the FORM and LEMMA columns, of which the former gets surface lexical forms of the current stage of derivation, and the latter gets the FORM data of the previous IG.

3.2 Coordination Structures

Among the most controversial annotation schemes are those of coordination structures (CS), which are groups of two or more tokens that are in coordination with each other, usually joined with conjunctions or punctuation, such as an *"and"* relation. The elements in coordination are the *conjuncts* of the CS, all of which are semantically linked to a single external head. A large variety of annotation methods are employed by different corpora, as thoroughly explained in (Popel et al., 2013). We chose three schemes to compare for our parser, which are illustrated in Figure 2. There does not seem to be a standard annotation rising as the best scheme, which is convenient because different schemes would have advantages and disadvantages against different formalisms and algorithms.

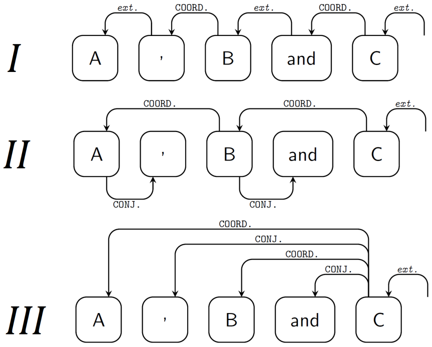


Figure 2: I) The original annotation scheme in the Turkish Treebank. II) *Swedish Style*, an alternative scheme in the manner of Talbanken (Nivre et al., 2006b). III) *Stanford Style*, another alternative scheme in the manner of the Stanford dependencies (De Marneffe and Manning, 2008), all with a head-right configuration as per (Popel et al., 2013), as would be appropriate for the predominantly head-final Turkish.

3.2.1 Current Coordination Representation

In the original Turkish Treebank, CSs are annotated as shown in scheme I in Figure 2, which appears to be problematic in several ways. This structure requires a prior conjunct to be connected to an intermediate conjunction, which in turn would be connected to a posterior conjunct, completing the coordination. The CS is then represented by the posterior conjunct, and the dependency relation between the prior conjunct and the conjunction must be identical to the dependency relation between the posterior conjunct and the external head, even if it would not semantically make sense.

Considering the tokens are processed incrementally from left to right during parsing, one difficulty with this method lies in correctly guessing the dependency relation between the prior argument and the conjunction before the posterior argument and the external head are even encountered, and unsurprisingly, directional parsers fail at this task more often than usual, resulting in added recall error for many dependency relations not necessarily related to coordinations. Another problem is that the scheme requires an intermediate conjunction or punctuation to work, which cannot be relied on even for edited texts, and would fare much worse if applied on web data. One final drawback of this method is that it is arguably more confusing for human annotators compared to a straightforward method in which the arguments in coordination are directly connected.

3.3 Proposed Coordination Representation

The drawbacks we have identified in the original CS annotation scheme encourage us to explore alternative approaches to coordinations. After investigating many annotation methods, we expect that the representation shown as the *Swedish Style* in Figure 2 will have the best performance in alleviating the issues described in Section 3.2.1.

Evaluating the *Swedish Style* representation, we observe that the CS does not depend on correctly placed conjunctions between the arguments, which increases compatibility in the absence of well-formatted sentences. Additionally, the dependency relation between the CS and the external head is not duplicated with this method, which should contribute to the reduction of recall error for many dependency types. Finally, we believe this scheme is easier for human annotators to understand and apply, and decreases the risk of annotation errors, which are very common in the Turkish Treebank.

4 Experiments

In order to practically evaluate our proposed IG and coordination representations, we first took our initial data set as our baseline, and then applied certain manual and automatic transformations to the data in order to create the experimental data sets. Since all of our data were based on a training corpus without an exclusive validation set, we decided to apply 10-fold cross-validation on all of our models to better evaluate the results.

For our tests on IG representations, we attempted to automatically transform our baseline corpus by populating the new `IG` column with boolean data derived from the IG relations in the gold-standard, and then automatically fill out the null lexical fields by an automatic morphological synthesis procedure using our morphological tool (Oflazer, 1994). The synthesis procedure, albeit a non-trivial implementation, successfully covered the majority (over 95%) of the lexical data, and we were able to manually annotate the remaining unrecognized tokens. To allow MaltParser to recognize the new fields, the CoNLL-X sentence format has been slightly adjusted and submitted as a custom input data format, and the baseline feature vector has been augmented with two extra features for the IG column information from the tokens on top of the Stack and Input pipes. The final model is named the *LexedIG* model.

On the other hand, we needed to perform a complete selective manual review of the corpus and correct numerous annotation errors in CSs before a healthy conversion could be made. Afterwards, we ran automatic conversion routines to map all CSs to the aforementioned *Swedish Style* and the commonly used *Stanford Style* in order to compare their specific performances. Since a sizeable amount of manual corrections were made before the conversions, we took the manually reviewed version as an intermediate model in order to distinguish the contribution of the automatic conversions from the manual review.

4.1 Metrics

For every model we evaluated via cross-validation, we made specific accuracy analyses and report the precision (P), recall (R) and F scores per dependency relation. Furthermore, we also calculated general accuracies as micro-averages from the cross-validation sets, for which we used two metrics, namely the labelled attachment score AS_L and the unlabelled attachment score AS_U, which are both accuracy metrics that compute the percentage of correctly parsed dependencies over all tokens, where the unlabelled metric only requires a match with the correct head, and the labelled metric additionally requires the correct dependency relation to be chosen.

	Baseline			LexedIG		
	P	R	F	P	R	F
ABLAT	61,46%	77,44%	0,69	61,50%	76,67%	0,68
APPOS	66,67%	12,87%	0,22	58,97%	11,39%	0,19
CLASS	72,98%	71,80%	0,72	72,57%	71,61%	0,72
COORD	83,95%	53,70%	0,66	83,57%	54,87%	0,66
DATIV	60,69%	71,57%	0,66	61,08%	70,68%	0,66
DERIV	**100,00%**	**100,00%**	**1,00**	**100,00%**	**100,00%**	**1,00**
DETER	91,18%	93,70%	0,92	91,23%	93,80%	0,92
INSTR	44,64%	38,38%	0,41	45,87%	40,96%	0,43
INTEN	87,99%	81,95%	0,85	87,35%	81,84%	0,85
LOCAT	73,40%	79,25%	0,76	73,90%	79,60%	0,77
MODIF	86,04%	81,58%	0,84	86,33%	81,74%	0,84
MWE	71,72%	58,72%	0,65	71,50%	59,42%	0,65
NEGAT	92,56%	70,00%	0,80	92,86%	73,13%	0,82
OBJEC	77,90%	71,36%	0,74	78,32%	71,92%	0,75
POSSE	87,44%	80,80%	0,84	86,58%	81,27%	0,84
QUEST	86,10%	77,16%	0,81	85,77%	77,16%	0,81
RELAT	70,00%	49,41%	0,58	70,49%	50,59%	0,59
ROOT	68,83%	99,77%	0,81	69,63%	99,77%	0,82
S.MOD	54,25%	50,25%	0,52	54,29%	50,92%	0,53
SENTE	93,25%	89,63%	0,91	93,20%	89,68%	0,91
SUBJE	69,54%	68,94%	0,69	69,87%	69,65%	0,70
VOCAT	69,61%	29,46%	0,41	69,23%	29,88%	0,42

Table 1: Specific accuracies per dependency relation for the IG-related models.

4.2 Results and Discussion

Our test results with the *LexedIG* model suggest that our proposed IG representation works perfectly well, as the perfect precision and recall scores of the original model for `DERIV` relations are preserved in the new model. Besides this, the reconstructed lexical information that we had populated the new model with caused only slight changes in overall accuracy that are not statistically significant, which is likely due to the sparsity of lexical data. Regardless, a model with lexical information for all tokens is essentially superior to a similarly performing model without such information. We foresee that being able to see lexical forms in the data would increase both the speed and the accuracy of human annotation. Additionally, as these experiments were done in preparation for the parsing of web data, we believe that in the near future, with the ability to unsupervisedly parse large amounts of data found on

	Baseline			Corrected			Swedish Style			Stanford Style		
	P	R	F	P	R	F	P	R	F	P	R	F
ABLAT	61,46%	77,44%	0,69	61,70%	79,46%	0,69	61,25%	79,19%	0,69	61,84%	80,20%	0,70
APPOS	66,67%	12,87%	0,22	62,86%	9,78%	0,17	65,79%	12,82%	0,21	67,57%	12,82%	0,22
CLASS	72,98%	71,80%	0,72	72,54%	71,76%	0,72	72,33%	74,52%	0,73	72,78%	74,22%	0,73
CONJU	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	79,78%	72,38%	0,76	76,99%	60,85%	0,68
COORD	83,95%	**53,70**%	0,66	83,88%	**54,23**%	0,66	79,15%	**64,64**%	0,71	73,82%	**58,68**%	0,65
DATIV	60,69%	71,57%	0,66	61,57%	71,77%	0,66	60,62%	72,83%	0,66	61,36%	73,81%	0,67
DERIV	100,00%	100,00%	1,00	100,00%	100,00%	1,00	100,00%	100,00%	1,00	100,00%	100,00%	1,00
DETER	91,18%	93,70%	0,92	91,08%	93,74%	0,92	91,14%	94,28%	0,93	91,15%	93,97%	0,93
INSTR	44,64%	38,38%	0,41	46,72%	39,48%	0,43	46,05%	41,08%	0,43	45,25%	41,49%	0,43
INTEN	87,99%	81,95%	0,85	87,30%	82,71%	0,85	87,46%	82,47%	0,85	87,83%	82,26%	0,85
LOCAT	73,40%	79,25%	0,76	73,92%	79,35%	0,77	72,33%	78,91%	0,75	72,42%	79,73%	0,76
MODIF	86,04%	81,58%	0,84	85,80%	81,47%	0,84	85,80%	81,84%	0,84	85,83%	81,06%	0,83
MWE	71,72%	58,72%	0,65	72,46%	59,09%	0,65	74,11%	58,87%	0,66	72,55%	60,18%	0,66
NEGAT	92,56%	70,00%	0,80	92,68%	66,28%	0,77	92,91%	73,29%	0,82	92,00%	71,43%	0,80
OBJEC	77,90%	71,36%	0,74	77,61%	71,54%	0,74	78,42%	72,12%	0,75	78,67%	72,08%	0,75
POSSE	87,44%	80,80%	0,84	87,03%	80,68%	0,84	87,69%	83,37%	0,85	87,25%	82,89%	0,85
QUEST	86,10%	77,16%	0,81	86,15%	77,78%	0,82	86,15%	78,05%	0,82	86,15%	78,05%	0,82
RELAT	70,00%	49,41%	0,58	71,67%	49,43%	0,59	72,13%	50,57%	0,59	69,35%	49,43%	0,58
ROOT	68,83%	99,77%	0,81	68,84%	99,49%	0,81	70,41%	99,79%	0,83	66,28%	99,81%	0,80
S.MOD	54,25%	50,25%	0,52	51,31%	49,28%	0,50	53,55%	50,09%	0,52	53,88%	49,91%	0,52
SENTE	93,25%	89,63%	0,91	92,74%	89,02%	0,91	93,50%	88,80%	0,91	93,36%	88,90%	0,91
SUBJE	69,54%	68,94%	0,69	69,61%	68,14%	0,69	69,89%	69,70%	0,70	69,75%	69,60%	0,70
VOCAT	69,61%	29,46%	0,41	67,86%	24,78%	0,36	61,05%	25,66%	0,36	69,62%	24,34%	0,36

Table 2: Specific accuracies per dependency relation for the coordination-related models.

the web, sparse data will no longer be a significant problem, and lexical data will gain further value.

A comparison of the alternative CS models with the baseline suggests that, while the manual correction itself did not cause a noticeable change, the automatic conversion procedures that it made possible resulted in significant improvements. The *Swedish Style* and *Stanford Style* models fared slightly better in the accuracy of some dependency types commonly joined in CSs such as SUBJECT, OBJECT, and DATIVE, INSTRUMENTAL and ABLATIVE.ADJUNCTs, but not always enough to warrant statistical significance. Apart from those, the largest improvement is in the COORDINATION relation itself, which had a slight drop in precision for both final models (likely due to the increased average dependency distances) but at the great benefit of the recall increasing from 53.70% to 58.68% for the *Stanford Style* and 64.64% for the *Swedish Style*.

	AS_U	AS_L
Baseline	74.5% ± 0.2	65.0% ± 0.2
LexedIG	74.6% ± 0.1	65.1% ± 0.2
Baseline	74.5% ± 0.2	65.0% ± 0.2
Corrected	74.5% ± 0.1	65.0% ± 0.2
Swedish Style	74.5% ± 0.2	**65.6**% ± 0.2
Stanford Style	73.2% ± 0.2	64.1% ± 0.2

Table 3: General parsing accuracies for all models, including standard error.

5 Conclusion

In this paper, we proposed novel annotation schemes for Turkish morphosyntactic sub-word units and coordination structures that are superior to the Turkish Treebank representations in terms of ease of use, parsing performance and/or compatibility with sentences that are not well-formed. Our findings substantiate our thesis that annotation schemes have both room for improvement and a high impact potential on parsing performance. In the light of our results, we intend to sustain our research and draw better annotation schemes for other syntactic structures such as copulae and modifier sub-types to serve not only Turkish, but also other languages with rich morphology.

Acknowledgments

The authors would like to acknowledge that this work is part of a research project supported by ICT COST Action IC1207 and TÜBİTAK 1001 (Grant Number 112E276).

References

Cristina Bosco, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice dell'Orletta, Alessandro Lenci, Leonardo Lesmo, Giuseppe Attardi, Maria Simi, Alberto Lavelli, et al. 2010. Comparing the influence of different treebank annotations on dependency parsing. In *LREC*.

Adriane Boyd and Detmar Meurers. 2008. Revisiting the impact of different annotation schemes on pcfg parsing: A grammatical dependency evaluation. In *Proceedings of the Workshop on Parsing German*, pages 24–32. Association for Computational Linguistics.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics.

Gülşen Eryiğit, Tugay Ilbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (IWPT)*, pages 45–55, Dublin, Ireland, October. Association for Computational Linguistics.

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

Dilek Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2002. Statistical morphological disambiguation for agglutinative languages. *Journal of Computers and Humanities*, 36(4):381–410.

Ryan McDonald and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

Joakim Nivre, Johan Hall, Jens Nilsson, Gülşen Eryiğit, and Stetoslav Marinov. 2006a. Labeled pseudo-projective dependency parsing with support vector machines. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 221–225, New York, NY.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006b. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1392–1395.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Stetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering Journal*, 13(2):99–135.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies*, pages 149–160, Nancy.

Kemal Oflazer, Bilge Say, Dilek Z. Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 261–277. Kluwer, London.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Kemal Oflazer. 2003. Dependency parsing with an extended finite-state approach. *Computational Linguistics*, 29(4):515–544.

Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zdeněk Žabokrtský. 2013. Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527, Sofia, Bulgaria, August. Association for Computational Linguistics.

Reut Tsarfaty and Yoav Goldberg. 2008. Word-based or morpheme-based? annotation strategies for modern hebrew clitics. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.