A Corpus for Modeling Morpho-Syntactic Agreement in Arabic: Gender, Number and Rationality

Sarah Alkuhlani and **Nizar Habash**
Center for Computational Learning Systems
Columbia University
{salkuhlani,habash}@ccls.columbia.edu

Abstract

We present an enriched version of the Penn Arabic Treebank (Maamouri et al., 2004), where latent features necessary for modeling morpho-syntactic agreement in Arabic are manually annotated. We describe our process for efficient annotation, and present the first quantitative analysis of Arabic morpho-syntactic phenomena.

1 Introduction

Arabic morphology is complex, partly because of its richness, and partly because of its complex morpho-syntactic agreement rules which depend on features not necessarily expressed in word forms, such as lexical rationality and functional gender and number. In this paper, we present an enriched version of the Penn Arabic Treebank (PATB, part 3) (Maamouri et al., 2004) that we manually annotated for these features.[1] We describe a process for how to do the annotation efficiently; and furthermore, present the first quantitative analysis of morpho-syntactic phenomena in Arabic.

This resource is important for building computational models of Arabic morphology and syntax that account for morpho-syntactic agreement patterns. It has already been used to demonstrate added value for Arabic dependency parsing (Marton et al., 2011).

This paper is structured as follows: Sections 2 and 3 present relevant linguistic facts and related work, respectively. Section 4 describes our annotation process and Section 5 presents an analysis of the annotated corpus.

[1]The annotations are publicly available for research purposes. Please contact authors. The PATB must be acquired through the Linguistic Data Consortium (LDC): http://www.ldc.upenn.edu/.

2 Linguistic Facts

Arabic has a rich and complex morphology. In addition to being both templatic (root/pattern) and concatenative (stems/affixes/clitics), Arabic's optional diacritics add to the degree of word ambiguity (Habash, 2010). This paper focuses on two specific issues of Modern Standard Arabic (MSA) nominal morphology involving the features of gender and number only: the discrepancy between morphological form and function and the complex system of morpho-syntactic agreement.

2.1 Form and Function

Arabic nominals (*i.e. nouns, proper nouns and adjectives*) and verbs inflect for gender: masculine (M) and feminine (F), and for number: singular (S), dual (D) and plural (P). These features are typically realized using a small set of suffixes that uniquely convey gender and number combinations: $+\phi$ (MS), ة+ $+\hbar$[2] (FS), ان+ *+An* (MD), تان+ *+tAn* (FD), ون+ *+wn* (MP), and ات+ *+At* (FP).[3] For example, the adjective ماهر *mAhr* 'clever' has the following forms among others: ماهر *mAhr* (MS), ماهرة *mAhrħ* (FS), ماهرون *mAhrwn* (MP), and ماهرات *mAhrAt* (FP). For a sizable minority of words, these features are expressed templatically, i.e., through pattern change, coupled with some singular suffix. A typical example of this phenomenon is the class of *broken*

[2]Arabic transliteration is presented in the Habash-Soudi-Buckwalter (HSB) scheme (Habash et al., 2007): (in alphabetical order) *AbtθjHxdðrzsšSDTDˇςγfqklmnhwy* and the additional symbols: ' ء, Â أ, Ă إ, Ā آ, ŵ ؤ , ŷ ئ , ħ ة, ý ى.

[3]Some suffixes have case/state varying forms, e.g., ون+ *+wn* appears as ين+ *+yn* (accusative/genitive case) and و+ *+w* (nominative construct state).

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 357–362,
Portland, Oregon, June 19-24, 2011.

plurals. In such cases, the form of the morphology (singular suffix) is inconsistent with the word's functional number (plural). For example, the word كاتب *kAtb* (MS) 'writer/scribe' has two broken plurals: كتّاب *ktAb* ($\frac{MS}{MP}$)[4] and كتبة *ktbħ* ($\frac{FS}{MP}$). In addition to broken plurals, Arabic has a class of *broken feminines* in which the feminine singular form is derived templatically: e.g., the adjective 'red' أحمر *ÂHmr* ($\frac{MS}{MS}$) and حمراء *HmrA'* ($\frac{MS}{FS}$). Verbs and nominal duals do not display this discrepancy. Ad hoc cases of form-function discrepancy also exist, e.g., خليفة *xlyfħ* ($\frac{FS}{MS}$) 'caliph', حامل *HAml* ($\frac{MS}{FS}$) 'pregnant', and طريق *Tryq* 'road' which can be both M and F ($\frac{MS}{BS}$). Arabic also has some non-countable collective plurals that behave as singulars morpho-syntactically although they may translate to English as plurals, e.g., تمر *tmr* ($\frac{MS}{MS}$) 'palm dates'.

2.2 Morpho-syntactic Agreement

Arabic gender and number features participate in morpho-syntactic agreement within specific constructions such as nouns and their adjectives and verbs and their subjects. Arabic agreement rules are more complex than the simple matching rules found in languages such as French or Spanish (Holes, 2004; Habash, 2010).

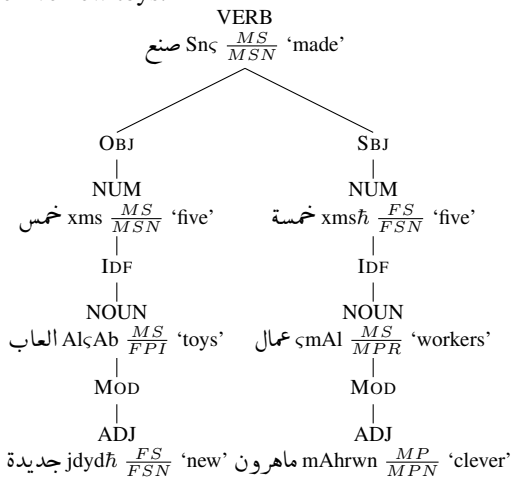
First, Arabic adjectives agree with the nouns they modify in gender and number except for plural irrational (non-human) nouns, which always take feminine singular adjectives. For example, the two plural words طالبات *TAlbAt* ($\frac{FP}{FPR}$)[5] 'students' and مكتبات *mktbAt* 'libraries' ($\frac{FP}{FPI}$) take the adjective 'new' as جديدات *jdydAt* ($\frac{FP}{FPN}$) and جديدة *jdydħ* ($\frac{FS}{FSN}$), respectively. Rationality is a morpho-lexical feature. There are nouns that are semantically rational/human but not morpho-syntactically, e.g., شعوب *šςwb* ($\frac{MS}{MPI}$) 'nations/peoples' takes a feminine singular adjective.[6]

Second, verbs and their nominal subjects have the same rules as nouns and their adjectives, except that, additionally, verbs in verb-subject (VSO) order only agree in gender and default to singular number. For example, the sentence 'the men traveled' can appear as الرجال سافروا *AlrjAl* ($\frac{MS}{MPR}$) *sAfrwA* ($\frac{MP}{MPN}$) or as سافر الرجال *sAfr* ($\frac{MS}{MSN}$) *AlrjAl* ($\frac{MS}{MPR}$).

Third, number quantification has unique rules (Dada, 2007), e.g., numbers over 10 always take a singular noun, while numbers 3 to 10 take a plural noun and *inversely* agree with the noun's functional gender.[7] Compare, for instance, خمس طالبات *xms* ($\frac{MS}{MSN}$) *TAlbAt* ($\frac{FP}{FPR}$) 'five [female] students' with خمسة طلاب *xmsħ* ($\frac{FS}{FSN}$) *TlAb* ($\frac{MS}{MPR}$) 'five [male] students' and خمسون طالبة *xmswn* ($\frac{MP}{BPN}$) *TAlbħ* ($\frac{FS}{FSR}$) 'lit. fifty [female] student[s]'. Figure 1 presents one example that combines the three phenomena mentioned above. The example is in a dependency representation based on the Columbia Arabic Treebank (CATIB) (Habash and Roth, 2009).

Finally, although the rules described above are generally followed, there are numerous exceptions that can typically be explained as some form of figure of speech involving elision or overridden rationality/irrationality. For example, the word جيش *jyš* ($\frac{MS}{MSI}$) 'army' can take the rational MP agreement in an elided reference to its members.

Figure 1: An example of a dependency tree with form-based and functional morphology features ($\frac{Form}{Function}$). صنع خمسة عمال ماهرون خمس العاب جديدة *Snς xmsħ ςmAl mAhrwn xms AlςAb jdydħ* 'Five clever workers made five new toys.'



[4]This nomenclature denotes ($\frac{Form}{Function}$).

[5]We specify rationality as part of the functional features of the word. The values of this feature are: rational (R), irrational (I), and not-applicable (N). N is assigned to verbs, adjectives, numbers and quantifiers.

[6]Rationality ('humanness' 'عاقل/غير عاقل') is narrower than animacy. English expresses it mainly in pronouns (*he/she* vs. *it*) and relativizers (*men who...* vs. *cars/cows which...*).

[7]*Reverse* gender agreement can be modeled as a form-function discrepancy, although it is typically not discussed as such in Arabic grammar.

3 Related Work

Much work has been done on Arabic computational morphology (Al-Sughaiyer and Al-Kharashi, 2004; Soudi et al., 2007; Habash, 2010). However, the bulk of this work does not address form-function discrepancy or morpho-syntactic agreement issues. This is unfortunately the case in some of the most commonly used resources for Arabic NLP: the Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004) and the Penn Arabic Tree Bank (PATB) (Maamouri et al., 2004). There are some important exceptions (Goweder et al., 2004; Smrž, 2007b; Elghamry et al., 2008; Abbès et al., 2004; Attia, 2008; Altantawy et al., 2010). We focus on comparing with two of these due to space restrictions.

Smrž (2007b)'s work contrasting illusory (form) features and functional features inspired our distinction of morphological form and function. However, unlike him, we do not distinguish between sub-functional (logical and formal) features. His ElixirFM analyzer (Smrž, 2007a) extends BAMA by including functional number and *some* functional gender information, but not rationality. This analyzer was used as part of the annotation of the Prague Arabic Dependency Treebank (PADT) (Smrž and Hajič, 2006). In the work presented here, we annotate for all three features completely in the PATB and we present a quantitative analysis of morpho-syntactic agreement patterns in it.

Elghamry et al. (2008) presented an automatic cue-based algorithm that uses bilingual and monolingual cues to build a web-extracted lexicon enriched with gender, number and rationality features. Their automatic technique achieves an F-score of 89.7% against a gold standard set. Unlike them, we annotate the PATB manually exploiting existing PATB information to help annotate efficiently and accurately.

4 Corpus Annotation

4.1 The Corpus

We annotated the Penn Arabic Treebank (PATB) part 3 (Maamouri et al., 2004) for functional gender, number and rationality. The corpus contains around 16.6K sentences and over 400K tokens.[8] All PATB tokens are already diacritized and lemma/part-of-speech (POS) disambiguated manually. Since verbs are regular in their form-to-function mapping, we annotate them automatically. Nominals account for almost half of all tokens ($\sim$ 197K tokens). The unique diacritized nominal types are almost 52K corresponding to 15,720 unique lemmas.

[8] All clitics are separated from words in the PATB except for the definite article +ال *Al*+.

4.2 Annotation Simplification

To simplify the annotation task, we made the following decisions. First, we decided to annotate nominals out of context except for the use of their lemmas and POS tags, which were already assigned manually in context in the PATB. The intuition here being that the functional features we are after are not contextually variable. We are consciously ignoring usage in figures-of-speech. Second, we normalized the case/state-variant forms of the number/gender suffixes and removed the definite article proclitic. The decision to normalize is conditioned on the manually annotated PATB POS tag. The normalized forms preserve the most important information for our task: the stem of the word and the number/gender suffix. These two decisions allow us exploit the PATB POS and lemma annotations to reduce the number of annotation decisions from 197K tokens and their lemmas to 21,148 morphologically normalized forms and 15,720 lemmas – an order of magnitude less decisions to make, which made the task more feasible both in terms of money and time. Of all nouns, adjectives and proper nouns, around 4.6% (tokens) and 27.2% (types) have no lemmas (annotated as DEFAULT, TBupdate, or nolemma). These cases make our out-of-context annotation very hard. We do not currently address this issue. A smaller set of closed class words (778 types corresponding to 35,675 tokens), e.g. pronouns and quantifiers, were annotated manually separately. The annotation speed averaged about 675 (words/lemmas) per hour.

4.3 Annotation Guidelines

We summarize the annotation guidelines here due to space restrictions. Full guidelines will be presented in a future publication. The core annotation task involves assigning the correct functional gender, number and rationality to nominals. Gender can be M, F, B (both), or U (unknown). Number can be S, D,

P, or U. And rationality can be R, I, B,[9] N or U. The annotators were given word clusters each of which consisting of a lemma and all of its simplified inflected forms appearing in the PATB. We also provided the POS and English gloss. Annotators were asked to assign the rationality feature to the lemma only; and the gender and number features to the inflected forms. Default form-based gender and number are provided. As for rationality, adjectives receive a default N and everything else gets I. The guidelines explained the form-function discrepancy problem, and the various morpho-syntactic agreement rules (Section 2) were given as tests to allow the annotators to make correct decisions. The issue figures-of-speech is highlighted as a challenge and annotators are asked to think of different contexts for the word in question.

4.4 Inter-Annotator Agreement

We computed inter-annotator agreement (IAA) over a random set of 397 lemma clusters with 509 word types corresponding to 4,781 tokens. The type-based IAA scores for words with known lemmas are 93.7%, 99.0% and 89.6% for gender, number and rationality respectively. The corresponding token-based IAA scores are 94.5%, 99.7% and 95.1%. The respective Kappa values (Cohen, 1960) for types are 0.87, 0.97, 0.82 and for tokens 0.89, 0.99, 0.92. Based on these scores, the number features is the easiest to annotate, followed by gender and rationality. This is explainable by the fact that number in Arabic is always expressed morphologically through affix or stem change, while gender is more lexical, and rationality is completely lexical. The corresponding IAA scores for all words (including words with unknown lemmas) drop to 86.8%, 94.9% and 82.9% (for types) and 93.5%, 99.2% and 94.0% (for tokens). The respective Kappa values for types are 0.74, 0.85, 0.73 and for tokens 0.87, 0.97, 0.90 The difference caused by missing lemmas highlights the need and value for complete annotations in the PATB. The overall high scores for IAA suggest that the task is not particularly hard for humans to perform, and that disambiguating information is crucial. Points of disagreement will be addressed in future extensions of the guidelines.

[9]The rationality value B is used for cases with lemma ambiguity, e.g., هيلتون *hyltwn* 'Hilton' can refer to the hotel chain or a member of the Hilton family.

5 Corpus Analysis

We present a quantitative analysis of the annotated corpus focusing on the issues that motivated it.

5.1 Form-Function Similarity Patterns

Table 1 summarizes the different combinations of form-function values of gender, number and rationality for nominals in our corpus. In terms of gender, the M value seems to be twice as common as F both in form and function. In 91.4% of all nominals, function and form agree. Adjectives show the most agreement (98.8%) followed by nouns (92.5%) and then proper nouns (74.6%). As for number, S is the dominant value in form (91.8%) and function (83.1%). Broken plurals ($\frac{S}{P}$) are almost 55% of all plurals. 99.5% of proper nouns are singular, which means that rationality is effectively irrelevant for proper nouns as a feature, since it is only relevant morpho-syntactically with plurals. Although the great majority of nouns are irrational, proper nouns tend to be almost equally split between rational and irrational. In terms of gender and number (jointly), 85% of all nominals have matching form and function values, with adjective having the highest ratio, followed by nouns and then proper nouns.

5.2 Morpho-syntactic Agreement Patterns

We focus on three agreement classes: Noun-Adj(ective), Verb-Subject (VSO and SVO orders) and Number-Noun (multiple configurations). We only consider structural bigrams in the CATIB (Habash and Roth, 2009) dependency version of the training portion of the PATB (part 3) used by Marton et al. (2011). See Figure 1 for an example. The total number of relevant bigrams is 39,561 or almost 11.6% of all bigrams. Over two-thirds are Noun-Adj, and around a quarter are Verb-Subject. For each agreement class, we compare using a simple agreement rule (parent and child values match) with using an implementation of the complex agreement rules summarized in Section 2. We also compare using form-based features or functional features.[10] Table 2 presents the percentage of bigrams we determine to *agree* (i.e. be grammatical) under different features and rules. Overall, simple (equality)

[10]Simple agreement between parent and child in gender *alone* is 83.2% (form) and 86.0% (function). The corresponding agreement for number is 82.0% (form) and 72.5% (function). The drop in the last number is due to broken plurals.

Feature	Values	Noun	Adjective	Proper	All
		69.2	18.2	12.5	100.0
GEN	M/M	64.5	48.9	71.3	62.5
	M/F	3.9	1.1	21.1	5.5
	M/B	0.4	0.0	3.4	0.7
	F/F	28.0	49.9	3.3	28.9
	F/M	3.1	0.1	0.8	2.3
	F/B	0.1	0.0	0.1	0.1
NUM	S/S	77.2	94.3	99.5	83.1
	S/P	12.2	1.5	0.4	8.7
	D/D	1.1	0.9	0.0	1.0
	P/P	9.5	3.3	0.1	7.2
RAT	-/I	94.7	—	45.3	71.2
	-/R	5.1	—	51.2	9.9
	-/B	0.3	—	3.5	0.6
	-/N	—	100.0	—	18.2
GEN+NUM	=/=	83.6	97.4	74.5	85.0

Table 1: Form-function discrepancy in nominals. All the numbers are percentages. Numbers in the first row are percentage of all nominals. Numbers in each column associated with a particular feature (or feature combination) and a particular POS are the percentage of occurrences within the POS. The second column specifies (Form/Function) values. =/= signifies complete match.

form-based gender and number agreement between parent and child is only 66.7%. Using functional values, the simple gender and number agreement moves only to 68.5%. Introducing complex agreement rules with form-based values (using the default N value for rationality of adjectives and I for other classes) increases grammaticality scores to 80.3% overall. However, with using both functional morphology features and complex agreement rules, the grammaticality score jumps to 93.6% overall. These results validate the need for both functional features and complex agreement rules in Arabic.

5.3 Manual Analysis of Agreement Problems

The cases we considered ungrammatical when applying complex agreement rules with functional features above add up to 2,540 instances. Out of these, we took a random sample of 423 cases and analyzed it manually. About 50% of all problems are the result of human annotation errors. Almost two-thirds of these errors involve incorrect rationality assignment and almost one-third involved incorrect gender. Incorrect number assignment occurs around 5% of the time. Treebank errors (as in POS or tree structure) are responsible for 20% of all agreement problems. Structure and POS tags are almost equal in their contribution. The rest of the agreement problems ($\sim$30%) are the result of special rules or figures-of-speech that are not handled. Figures of speech account for about 7% of all error cases (or less than 0.5% of all nominals). The most common cases of unhandled rules include not modeling conjunctions, which affect number agreement, followed by gender-number invariable forms of some adjectives. After this error analysis, we identified 379 lemmas involved in incorrect rationality-affected agreement (as per our rules). All of these cases had a PI features but did not agreed as FS. Out of these lemmas, 204 were corrected manually as R. The functional agreement with rules jumped from 93.6% to 95.7% (a 33% error reduction).

	Features × Agreement			
	Form-based		Functional	
Constructions	Simple	Rules	Simple	Rules
Noun-Adj (69.2)	66.7	81.7	69.2	94.8
Verb-Subj (26.7)	73.7	81.5	75.0	90.2
Num-Noun (4.0)	21.6	48.8	14.5	94.4
All (100.0)	66.7	80.3	68.5	93.6

Table 2: Analysis of gender+number agreement patterns in the annotated corpus. All numbers are percentages.

6 Conclusions and Future Work

We presented a large resource enriched with latent features necessary for modeling morpho-syntactic agreement in Arabic. In future work, we plan to use both corpus annotations and agreement rules to automatically learn functional features for unseen words and detect and correct annotation errors. We also plan to extend agreement rules to include complex structures beyond bigrams.

Acknowledgments

We would like to thank Mona Diab, Owen Rambow, Yuval Marton, Tim Buckwalter, Otakar Smrž, Reem Faraj, and May Ahmar for helpful discussions and feedback. We also would like to especially thank Ahmed El Kholy and Jamila El-Gizuli for help with the annotations. The first author was funded by a scholarship from the Saudi Arabian Ministry of Higher Education. The rest of the work was funded under DARPA projects number HR0011-08-C-0004 and HR0011-08-C-0110.

References

Ramzi Abbès, Joseph Dichy, and Mohamed Hassoun. 2004. The Architecture of a Standard Arabic Lexical Database. Some Figures, Ratios and Categories from the DIINAR.1 Source Program. In Ali Farghaly and Karine Megerdoomian, editors, *COLING 2004 Computational Approaches to Arabic Script-based Languages*, pages 15–22, Geneva, Switzerland, August 28th. COLING.

Imad Al-Sughaiyer and Ibrahim Al-Kharashi. 2004. Arabic Morphological Analysis Techniques: A Comprehensive Survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

Mohamed Altantawy, Nizar Habash, Owen Rambow, and Ibrahim Saleh. 2010. Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

Mohammed Attia. 2008. *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation.* Ph.D. thesis, The University of Manchester, Manchester, UK.

Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, University of Pennsylvania. LDC Cat alog No.: LDC2004L02, ISBN 1-58563-324-0.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Ali Dada. 2007. Implementation of the Arabic Numerals and their Syntax in GF. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 9–16, Prague, Czech Republic.

Khaled Elghamry, Rania Al-Sabbagh, and Nagwa El-Zeiny. 2008. Cue-based bootstrapping of Arabic semantic features. In *JADT 2008: 9es Journées internationales d'Analyse statistique des Données Textuelles*.

Abduelbaset Goweder, Massimo Poesio, Anne De Roeck, and Jeff Reynolds. 2004. Identifying Broken Plurals in Unvowelised Arabic Text. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 246–253, Barcelona, Spain, July.

Nizar Habash and Ryan Roth. 2009. CATiB: The Columbia Arabic Treebank. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 221–224, Suntec, Singapore.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Yuval Marton, Nizar Habash, and Owen Rambow. 2011. Improving Arabic Dependency Parsing with Form-based and Functional Morphological Features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, Oregon, USA.

Otakar Smrž and Jan Hajič. 2006. The Other Arabic Treebank: Prague Dependencies and Functions. In Ali Farghaly, editor, *Arabic Computational Linguistics: Current Implementations*. CSLI Publications.

Otakar Smrž. 2007a. ElixirFM – implementation of functional arabic morphology. In *ACL 2007 Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 1–8, Prague, Czech Republic. ACL.

Otakar Smrž. 2007b. *Functional Arabic Morphology. Formal System and Implementation*. Ph.D. thesis, Charles University in Prague, Prague, Czech Republic.

Abdelhadi Soudi, Antal van den Bosch, and Günter Neumann, editors. 2007. *Arabic Computational Morphology. Knowledge-based and Empirical Methods*, volume 38 of *Text, Speech and Language Technology*. Springer, August.