Поиск ошибок в корпусе с помощью MTE-разметки

Detecting errors in a corpus using mte-annotation

Копотев М. В. (mihail.kopotev@helsinki.fi)

Хельсинкский университет, Финляндия

В докладе описывается формат морфосинтаксического аннотирования MTE (MulText-East). На примере корпуса ХАНКО определяются возможности его применения для поиска ошибок морфологического аннотирования, выявляемых на основе анализа частоты совместной встречаемости граммем.

1. Описание формата

Формат MTE (MulText-East) — это многоязычный стандарт, разработанный для компактного и унифицированного представления морфосинтаксической информации для ряда европейских языков. Эта спецификация подготовлена группой из Любляны под руководством Т. Ержавеца для болгарского, чешского, эстонского, венгерского, румынского, словенского и английского языков (см. Erjavec [в печати]; Multext-East [электронный ресурс]). В 2007–2008 гг. стандарт был адаптирован к русскому языку (Sharoff et al., 2008). В настоящее время существует четвертая версия формата, продолжается работа по ее применению к другим славянским языкам (в том числе и другим восточнославянским).

Суть спецификации состоит в стандартном пословном морфосинтаксическом представлении в виде так называемых «тегсетов» (*tagsets*), или наборов морфологических показателей. Основная идея MTE заключается в том, что для каждой части речи представлена спецификация, устанавливающая набор атрибутов: **род, число, падеж** и т.д. с определенным списком значений (тегов): мужской, женский, средний. Место атрибута в тегсете фиксировано, его значение выражается буквой латинского алфавита. Первое место в тегсете всегда занимает показатель части речи, заданный заглавной буквой. Таким образом, морфосинтаксическое описание каждой текстоформы передается с помощью буквенного кода. Например,

Семеняка	Npfsny--	Семеняка
не	Qs	не
играла	V-is-sfa-p---	играть
в	Sps-	в
эти	Pd--paa--	этот
игры	Ncfpan--	игра
.	X	.

Для существительного (N) 1) собственного (p), 2) женского рода (f), 3) единственного числа (s), 4) в именительном падеже (n), 5) одушевленного (y) тегсет выглядит так:

Npfsny--

Если атрибуты по каким-то причинам не могут быть применены к определенной лексеме или текстоформе, место переменной заполняется дефисом. Например,

свято Afpns-s- святой

Прилагательное (A) качественно-относительное (f), в положительной степени (p), среднего рода (n), единственного числа (s) в краткой форме (s) не определено по атрибутам «падеж» (шестая позиция) и «сокращение» (восьмая позиция).

Спецификация MTE, представленная в работе (Sharoff et al. 2008), преследовала цели унифицировать формат для использования с разными языками и разными корпусами, что привело к уменьшению числа атрибутов. Следующая таблица дает представление о числе атрибутов для разных частей речи (см. табл. 1).

Два последних нуждаются в комментарии. Класс «Abbreviation» вызвал оживленную дискуссию при адаптации схему к русскому материалу, в результате которой было решено собрать в этот класс все аббревиатуры и сокращения, маркируя их морфологический тип (именные, наречные), а также род, число и падеж. Класс «Residual» зарезервирован для разного рода неопознанных случаев. Как видно из таблицы 1, число атрибутов варьируется от 0 (для «Residual») до 10 (для «Verb»). Таблица 2 дает представление о распределении атрибутов по частеречным классам.

Табл. 1. http://corpus.leeds.ac.uk/mocky/back.1_div.1.html)

Name (en)	Code (en)	Attributes
Noun	N	6
Verb	V	10
Adjective	A	6
Pronoun	P	7
Adverb	R	1
Adposition	S	3
Conjunction	C	4
Numeral	M	6
Particle	Q	1
Interjection	I	1
Abbreviation	Y	4
Residual	X	0

Табл 2. http://corpus.leeds.ac.uk/mocky/back.1_div.2.html)

Category (en)	Attribute (en)	Position
Abbreviation	Case	4
Abbreviation	Gender	2
Abbreviation	Number	3
Abbreviation	Syntactic_Type	1
Adjective	Case	5
Adjective	Definiteness	6
Adjective	Degree	2
Adjective	Gender	3
Adjective	Number	4
Adjective	Type	1
Adposition	Case	3
Adposition	Formation	2
Adposition	Type	1
Adverb	Degree	1
Conjunction	Coord_Type	3
Conjunction	Formation	2
Conjunction	Sub_Type	4
Conjunction	Type	1
Interjection	Formation	2
Noun	Animate	5
Noun	Case	4
Noun	Case2	6
Noun	Gender	2
Noun	Number	3
Noun	Type	1
Numeral	Animate	6
Numeral	Case	4
Numeral	Form	5
Numeral	Gender	2
Numeral	Number	3
Numeral	Type	1
Particle	Formation	1
Pronoun	Animate	7
Pronoun	Case	5
Pronoun	Gender	3
Pronoun	Number	4
Pronoun	Person	2
Pronoun	Syntactic_Type	6
Pronoun	Type	1
Verb	Aspect	9
Verb	Case	10
Verb	Definiteness	8
Verb	Gender	6
Verb	Number	5
Verb	Person	4
Verb	Tense	3
Verb	Type	1
Verb	VForm	2
Verb	Voice	7

Для корпуса ХАНКО эта спецификация была незначительно изменена в силу более детального представления морфологии в этом корпусе. Самым существенным изменением стало исключение класса «Abbreviation». Однако соответствующие атрибуты добавляются в другие части речи: акронимы для существительных (*США*) и сокращения для существительных (*г*[*од*]), глаголов (*см.*[*отри*]), прилагательных (*проч.*[*ее*]), местоимений (до *н.*[*ашей*] эры) и числительных (*тыс.*[*яча*]).. Кроме этого, были добавлены некоторые атрибуты, например, Pluralia tantum, дробные числительные и условное наклонение. Подробную информацию о ХАНКО в формате MTE можно найти по адресу **www.ling.helsinki.fi/projects/hanco/mte**.

2. Использование формата MTE для поиска ошибок аннотирования

Формат MTE обладает рядом особенностей, делающих его перспективным инструментом анализа собственно морфологических явлений и оценки аккуратности разметки. Преимуществом такого формата являются следующее.

1. **Простота и компактность**. Корпус в формате MTE представляет собой простой текстовый файл и в этом виде не требует специальных программ и оболочек. Для поиска и обработки данных можно использовать стандартные средства (например, команды семейства grep в Unix).
2. **Кроссязычность**. Стандарт MTE изначально задумывался таким образом, чтобы варианты тегсетов для разных языков были максимально близки с формальной стороны. Конечно, это невозможно соблюсти в полной мере (ср., например, количество падежей в разных языках или наличие полной формы прилагательного в русском языке). Однако в целом позиция

сходных атрибутов в тегсете и теги для совпадающих значений одинаковы для всех языков (например, **n** в пятой позиции для номинатива существительных). Эта особенность дает возможность проводить корпусные исследования по сопоставительной морфологии.

3. **Частотный анализ тегсетов**. Формат MTE позволяет анализировать не только частотность отдельных граммем, как это сделано в работах (Josselson 1953; Greenberg 1974; Копотев 2008; серия статей в (Корпусные исследования 2009) и др.), но и сочетаний граммем: целых тегсетов или их фрагментов[1].

Однако кроме этих особенностей формат MTE позволяет использовать его еще для одной процедуры, а именно для поиска ошибок в корпусе[2]. Рабочая гипотеза, которая лежит в основе этой процедуры, состоит в следующем:

> тегсеты с низкой частотой содержат большее количество ошибок, чем тегсеты с высокой частотой, поскольку хвосты частотного распределения могут быть вызваны конфликтным сочетанием тегов.

Конечно, это не значит, что все уникальные тегсеты ошибочны. Кроме того, метод не может уловить все ошибки в корпусе, поскольку теггер, с помощью которого размечался корпус, мог быть неправильно настроен и выдавать, соответственно большой массив неправильных аннотаций (известный пример такого рода — лексема *Пу́тина,* разобранная как *пути́на*). Еще одно ограничение связано с зависимым приписыванием тегов. Так, признаки падежа и числа для существительного, извлеченные из флексии, очевидно, не будут конфликты, хотя и могут быть ошибочными (например, *Людмила Пу́тина*, разобранная как форма Род. пад. Ед. ч.). Представляется, что предложенный метод особенно эффективен при оценке качества ручной обработки корпуса после автоматического аннотирования. С помощью этого метода можно частично исправить или хотя бы оценить степень аккуратности аннотирования. Для настоящего доклада была проанализирована верность/ошибочность всех тегсетов с низкой частотой для глагола, существительного и прилагательного в корпусе ХАНКО. Подсчет проводился до тех пор, пока все тегсеты, входящие в ранг, не оказывались правильными. Для существительных и глаголов это тегсеты с частотой 3, для прилагательных — 4.

2.1. Глаголы

На графике 1. показано распределение всех глагольных тегсетов.

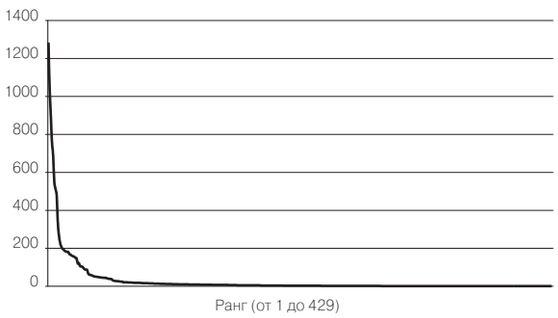

График 1. Глагол. Распределение тегсетов

Зона хвостов (с частотностью менее 10) отдельно показана на графике 2, на котором можно видеть длинный ряд тегсетов, имеющих последний ранг.

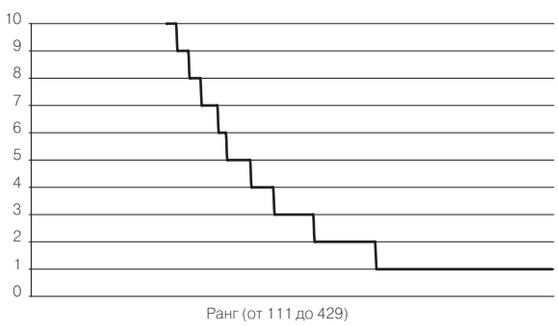

График 2. Глагол. Распределение низкочастотных тегсетов

Всего в корпусе встретилось 145 глагольных тегсетов с частотой 1. Из них 82 (56.6 %) оказались ошибочными. Приведу примеры.

1. замирившимися V-ps-**pm**afeir- замириться
   Ошибочно определен род причастия в форме множественного числа.
2. посрывало V-**-**-s3sna-e-u- посрывать
   Не определены форма и время глагола (вторая и третья позиции).
3. приводится V-ip-s-p-p-u- приводить
   Пропущен тег лица (пятая позиция).

Любопытны также и правильно определенные тегсеты. Часто достаточно одного редко встречающегося морфологического параметра, чтобы весь тегсет стал уникальным. Так, в ХАНКО последовательно маркируются двувидовые глаголы. Посколь-

[1] Близкий к этому подход разрабатывается в (Arppe 2001, Janda & Lyashevskaya [в печати]), которые используют термин «grammatical profile» для характеристики частотного распределения морфологических форм лексемы.

[2] Другие методы оценки корпуса и поиска ошибок предложены в (Brants 1995; Dickinson & Meurers 2003; Pîrvan & Tufiş 2006 и особенно Dickinson 2003).

ку они в целом достаточно редки, количество тегсетов двувидовых глаголов с частотой 1 довольно значительно.

4. использовалось V-is-snp-**b**-u- использовать
5. регламентирующих V-pp-p-af**b**g-- регламентировать

Еще одним атрибутом, влияющим на частотность всего тегсета, является сокращение (s в 13 позиции):

6. Н. V-p--p-pf-gu**s** называемый

Среди глагольных тегсетов, встретившихся в корпусе два раза, количество ошибок резко падает и составляет всего 2 %. Среди тегсетов с частотой 3 ошибочных не встретилось совсем. График 3 показывает соотношение количества ошибочных чтений для низкочастотной глаголов.

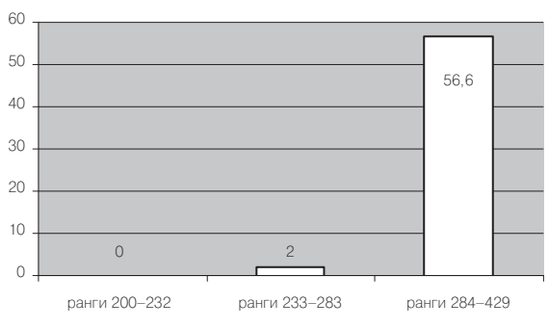

График 3. Глагол. % ошибочных тегсетов

2.2. Существительные

На графике 4 показано распределение тегсетов существительных, встретившихся в корпусе 10 и менее раз.

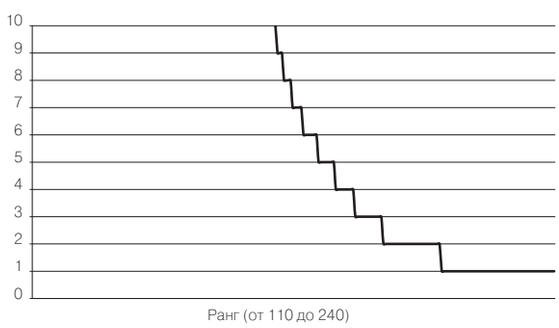

График 4. Существительное. Распределение низкочастотных тегсетов

В корпусе зафиксировано 52 тега с частотой 1, из которых 20 ошибочных (38,46 %).

Это ошибки такого рода.

7. ВЭФ N-----а ВЭФ
   Часть тегов не отмечено.
8. интернет-компаниях Nffpl--- интернет-компания
   Смешана разметка существительного (компания) и прилагательного (интернет-).
9. Лефортовской **N**-fsl--- Лефортовский
   Неверно определена часть речи.
10. Абрамович N-m-ny-- Абрамович
    Часть тегов не отмечено.

Самыми частыми атрибутами, приводящими к резкому снижению частотности тегсета, являются теги, маркирующие сокращение и акронимы:

11. ГЭС Ncfsin-**a** ГЭС
12. ул. Ncfsnn-**s** улица

а также сочетание тегов «собственное» (**p** во второй позиции) и «множественное число» (**p** в четвертой позиции):

13. Сезары N**p**m**p**an-- Сезар
14. Людвигов N**p**m**p**gn-- Людвиг

График 5 показывает соотношение ошибочных тегсетов в трех самых низкочастотных группах. Из него видно, что тегсеты с частотой 3 не содержат ошибок.

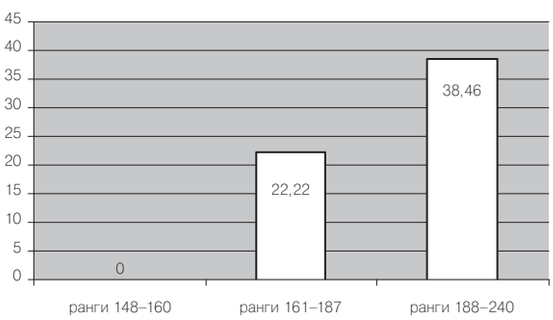

График 5. Существительное. % ошибочных тегсетов

2.3. Прилагательные

На графике 6 показано распределение тегсетов прилагательных с частотой меньше 10.

Всего в корпусе встретилось 50 уникальных тегсетов для прилагательных, из которых ровно половина оказалась ошибочной. Часть из них — по причине неверного сочетания тегов:

15. дедушкиных As-**fp**gf- дедушкин
    Сочетание женского рода и множественного числа.

В значительной части ошибочных тегсетов не содержится тега, определяющего тип прилагательного:

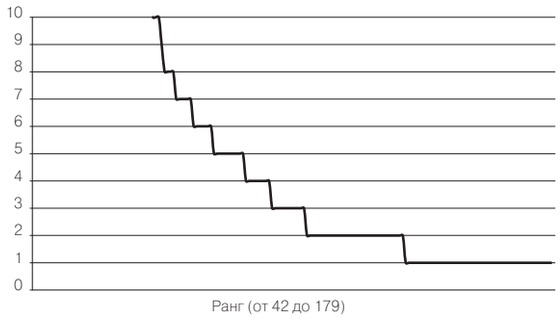

График 6. Прилагательное. Распределение низкочастотных тегсетов

16. интерросовского A--msg-- интерросовский
    Пропущен тег f во второй позиции.

Еще одна часть ошибочных чтений не содержит тега полной/краткой формы:

17. лучшему Afsnsd-- хороший
    Пропущен тег f в седьмой позиции.

Если говорить о правильных, но редко встречающихся тегсетах, то, как и для существительного и глагола, они часто вызваны сокращением слова:

18. Солнечн. Afpfsnfs солнечный
    Тег s в девятой позиции.

Ожидаемо редки притяжательные прилагательные, что приводит к снижению частотности всего тегсета.

19. Божим As-msi-- Божий
    Тег s во второй позиции.

На графике 7, как и в предыдущих случаях, показана доля ошибочно размеченных тегсетов в самых низкочастотных случаях.

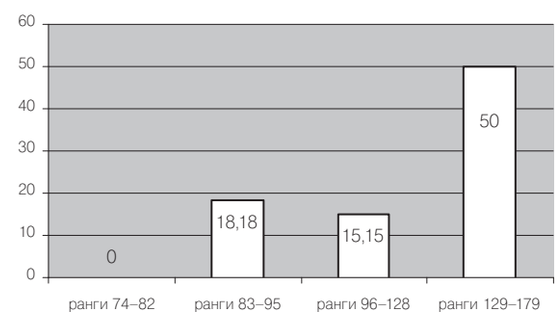

График 7. Прилагательное. % ошибочных тегсетов

На графике видно, что для прилагательных картина распределения ошибок несколько иная: большой процент содержат не только две самых низкочастотных группы (частотностью 1 и 2), но и третья (частотность 3, ранги 83–95). Большинство из этих ошибок связано с неверным сочетанием тегов мн. числа и рода. Лишь четвертая с конца группа не содержит ошибок.

3. Выводы

График 8 дает общее представление о количестве тегсетов для трех рассмотренных частей речи (в скобках приведены абсолютные данные). Он показывает совокупную долю всех найденных ошибок к общему количеству тегсетов той или иной части речи.

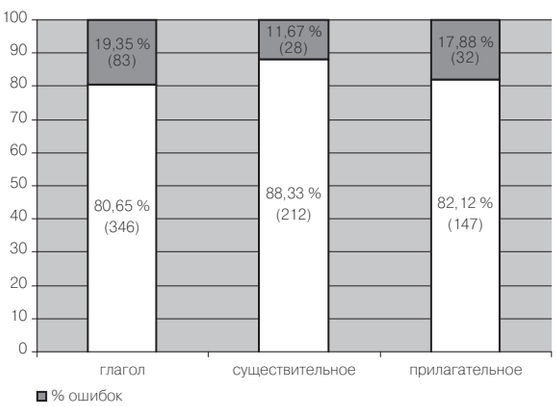

График 8. Совокупный % ошибочных тегсетов

Представляется, что анализ низкочастотных тегсетов является полезным методом поиска ошибок, оценки и улучшения морфологической разметки русского корпуса. На примере трех частей речи удалось показать, что в зону хвостов попадает непропорционально много ошибочных тегсетов. Представляется, что на более объемном корпусе эти результаты будут еще более контрастными, поскольку частотность правильных, но редких тегсетов увеличится. При этом надо учитывать, что на частотность тегсетов влияет и схема аннотирования: чем она грубее, тем легче ее применить и тем меньше правильных низкочастотных тегсетов окажется в корпусе. Наконец, надо заметить, что соотношение ранга и частотности, установленное в свое время Джорджем Ципфом для распределения лексем, в целом верно и для морфосинтаксических тегсетов, хотя эта параллель еще требует дальнейшего уточнения и обоснования.

Литература

1. *Arppe A.* Focal points in frequency profiles — how some word forms in a paradigm are more significant than others in Finnish // Proceedings of the 6th Conference on Computational Lexicography and Corpus Research, June 28–30, 2001, University of Birmingham, Birmingham, 2001.
2. *Brants Th.* Tagset Reduction Without Information Loss // Proceedings of the 33rd Annual Meeting of the ACL. Cambridge, MA. 1995. P. 287–289.
3. *Dickinson M, Meurers, D.* Detecting Errors in Part-of-Speech Annotation // Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03). Budapest, Hungary, 2003.
4. *Dickinson, M.* Error detection and correction in annotated corpora. PhD thesis. The Ohio State University, 2003. Доступно по адресу: etd.ohiolink.edu/view.cgi?osu1123788552.
5. *Erjavec. T.* MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. // Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2010), Malta. [В печати].
6. *Greenberg J. H.* The relation of frequency to semantic feature in a case language (Russian) // K. Denning and S. Kemmer (eds), On Language, Selected Writings of Joseph H. Greenberg, Stanford, 1990 (1974). P. 207–226.
7. *Janda L., Lyashevskaya O.* Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian [в печати].
8. *Josselson H. H.* Подсчет ходовых слов русского языка, Detroit (MI), 1953.
9. *Multext-East* Home Page. [Интернет-ресурс. Доступен по адресу: nl.ijs.si/ME]
10. *Pîrvan, F. Tufiş. D.* Tagsets Mapping and Statistical Training Data Cleaning-up // Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, May 2006. Genoa. P. 385–390.
11. *Sharoff, S, Kopotev, M., Erjavec, T, Feldman A., Divjak, D.* Designing and evaluating Russian tagset // Proceedings Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, May, 2008. Доступен по адресу: www.lrec-conf.org/proceedings/lrec2008/summaries/78.html.
12. *Корпусные* исследования по русской грамматике, Москва: Пробел-2000, 2009.