

## СПОСОБЫ ПРЕДСТАВЛЕНИЯ МИ, ИСПОЛЬЗУЕМЫЕ РАЗЛИЧНЫМИ СИСТЕМАМИ КОМПЬЮТЕРНОГО МА

В данной работе будут рассмотрены способы представления МИ, используемые различными системами компьютерного МА, а также описаны правила определения подчинительной связи между словами.

Один из вопросов, требующих особого рассмотрения, – способ представления МИ. Задание МИ должно быть компактно, позволять быстро и просто извлекать отдельные морфологические характеристики слова (например, только род или только часть речи).

Обычно постоянная и переменная морфологические информации хранятся в виде строк текста. Иногда ее представляют в виде фрейма, в котором для хранения каждой из характеристик отведено отдельное поле. Для конкретной словоформы заполняются не все поля, а только те которые для нее характерны.

В первом и во втором случае сталкиваемся с резервированием лишнего места и необходимостью использования сложных процедур интерпретации.

Третий способ хранения МИ – в виде битовых полей. Он сочетает в себе удобства обработки и экономное хранение МИ, поэтому более предпочтителен. Рассмотрим способы задания морфологической информации с помощью битовых полей, примененные в модуле морфологического анализа LINGUIST, в модуле лемматизации, разработанном в отделе РРО ИПИИ.

### **Формат представления морфологической информации модулем морфологического анализа LINGUIST**

МИ слова в модуле LINGUIST полностью описывается структурой TGramInfo. Эта структура содержит три поля, первое из которых (wInfo) относится к самому слову – описывает постоянную МИ, а два других (gInfo & Flags) описывают его форму – переменная МИ:

```
typedef struct _GramInfo
{
    WORD16 wInfo;
    WORD16 gInfo;
    WORD16 Flags;
} TGramInfo;
```

Первое поле (wInfo), а точнее, младшие его 6 бит, содержат информацию о типе слова (части речи) по классификации LINGUIST, приведенной в таблице 1. Второе поле структуры TGramInfo - gInfo - содержит информацию о форме слова (такую, как падеж или род, то есть переменная МИ), приведенную в таблице 2.

***Вопрос: Как организована морфологическая база?***

Таблица 1 – Описание постоянной МИ в модуле LINGUIST

| wInfo & 0x3F | Часть речи                                                  |
|--------------|-------------------------------------------------------------|
| 0x01         | Глагол несовершенного вида                                  |
| 0x02         | Непереходный глагол несовершенного вида                     |
| 0x03         | Глагол совершенного вида                                    |
| 0x04         | Непереходный глагол совершенного вида                       |
| 0x05         | Двувидовой глагол                                           |
| 0x06         | Непереходный двувидовой глагол                              |
| 0x07         | Неодушевленное существительное мужского рода                |
| 0x08         | Одушевленное существительное мужского рода                  |
| 0x09         | Одушевленное - неодушевленное существительное мужского рода |
| 0x0A         | Неодушевленное существительное мужского рода                |
| 0x0B..0x0C   | Одушевленное существительное мужского рода                  |
| 0x0D         | Неодушевленное существительное женского рода                |
| 0x0E         | Одушевленное существительное женского рода                  |
| 0x0F         | Одушевленное - неодушевленное существительное женского рода |
| 0x10         | Неодушевленное существительное среднего рода                |
| 0x11         | Одушевленное существительное среднего рода                  |
| 0x12         | Одушевленное - неодушевленное существительное среднего рода |
| 0x13         | Неодушевленное существительное общего рода                  |
| 0x14         | Одушевленное существительное общего рода                    |
| 0x15         | Неодушевленное существительное мужского/среднего рода       |
| 0x16         | Одушевленное существительное мужского/среднего рода         |
| 0x17         | Неодушевленное существительное женского/среднего рода       |
| 0x18         | Неодушевленное существительное множественного числа         |
| 0x19..0x1A   | Прилагательные                                              |
| 0x1B         | Притяжательные местоимения                                  |
| 0x1C         | Местоименные прилагательные                                 |
| 0x1D..0x20   | Местоимения                                                 |
| 0x21..0x22   | Числительное                                                |
| 0x23         | Собирательное числительное                                  |
| 0x24         | Порядковое числительное                                     |
| 0x25..0x27   | Имена собственные                                           |
| 0x28..0x29   | Отчества                                                    |
| 0x2A         | Фамилии                                                     |
| 0x2B..0x2F   | Географические названия                                     |
| 0x30         | Вводное слово                                               |
| 0x31         | Междометие                                                  |
| 0x32         | Предикатив                                                  |
| 0x33         | Предлог                                                     |
| 0x34         | Союз                                                        |
| 0x35         | Частица                                                     |
| 0x36         | Наречие                                                     |
| 0x37         | Сокращенное существительное                                 |
| 0x38         | Сокращенное прилагательное                                  |
| 0x39         | Сокращенное вводное слово                                   |
| 0x3A         | Обособленная сравнительная степень                          |
| 0x3B..0x3C   | Аббревиатура                                                |

Таблица 2 – Описание переменной МИ в модуле LINGUIST

| Бит   | Значение          | Значение                                                                                                                                                                                                                |
|-------|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 0-2   | 0x0001            | инфинитив                                                                                                                                                                                                               |
|       | 0x0002            | Повелительное наклонение                                                                                                                                                                                                |
|       | 0x0003            | будущее время                                                                                                                                                                                                           |
|       | 0x0004            | настоящее время                                                                                                                                                                                                         |
|       | 0x0005            | прошедшее время.                                                                                                                                                                                                        |
| 3-4   | 0x0008            | первое лицо                                                                                                                                                                                                             |
|       | 0x0010            | второе лицо                                                                                                                                                                                                             |
|       | 0x0018            | третье лицо.                                                                                                                                                                                                            |
| 5-6   | 0x0000            | личная форма глагола                                                                                                                                                                                                    |
|       | 0x0020            | действительное причастие                                                                                                                                                                                                |
|       | 0x0040            | страдательное причастие                                                                                                                                                                                                 |
|       | 0x0060            | деепричастие.                                                                                                                                                                                                           |
| 7     | 0x0080            | Сравнительная степень прилагательных.                                                                                                                                                                                   |
| 8     | 0x0100            | Краткая форма прилагательных и страдательных причастий.                                                                                                                                                                 |
| 9-10  | 0x0200            | мужской род, единственное число                                                                                                                                                                                         |
|       | 0x0400            | женский род, единственное число                                                                                                                                                                                         |
|       | 0x0600            | средний род, единственное число                                                                                                                                                                                         |
| 11    | 0x0800            | Множественное число склоняющихся слов.                                                                                                                                                                                  |
| 12-14 | 0x1000-<br>0x7000 | Падеж. Может иметь значение от 0 до 7, что соответствует восьми падежам – именительному, родительному, дательному, винительному, творительному, предложному, второму родительному и второму предложному соответственно. |
| 15    | 0x8000            | Возвратная форма. Бит выставлен у возвратных форм глаголов и некоторых прилагательных и произошедших от них существительных.                                                                                            |

Таким образом, например, действительному причастию прошедшего времени творительного падежа множественного числа в возвратной форме, построенному от глагола "делать" ("делавшимися") будет соответствовать значение gInfo 0xC825.

Третье поле структуры - Flags – содержит дополнительную грамматическую информацию о форме слова, которая не уместилась в gInfo – одушевленность, а также некоторые дополнительные пометы, о которых будет сказано ниже. Признак одушевленности анализируют лишь для винительного падежа прилагательных и причастий, так как в остальных случаях словоформы графически совпадают. Если форма слова соответствует винительному падежу одушевленного, в Flags выставлен бит 0x0001, если неодушевленного - 0x0002.

При работе с морфологической информацией модуля LINGUIST используют приведенные ниже макроопределения, извлекающие некоторые часто используемые характеристики.

*Пример 1.* Падеж склоняющегося слова:

```
#define GetCase( gInfo ) ( ( (gInfo) & 0x7000 ) >> 12 )
```

Макроопределение извлекает падеж склоняющегося слова. Именительному падежу соответствует значение 0, предложному, соответственно, 5. Значения 6 и 7 соответствуют так называемым второму родительному и второму предложному падежам.

*Пример 2.* Род:

```
#define GetGender( gInfo ) ( ( (gInfo) & 0x0600 ) >> 9 )
```

Макроопределение извлекает род. Мужскому роду соответствует значение 1, женскому – 2, среднему - 3.

### **Представление морфологической информации в модуле лемматизации отдела РРО ИПИИ**

Способ представления МИ в модуле лемматизации РРО представлен в таблице 3. Он позволяет экономить 2 байта на каждой записи, по сравнению с модулем LINGUIST. Это является существенной экономией дискового пространства, так как предполагаемое количество записей – 2 млн. Кроме того, определение значений отдельных категорий МИ в модуле лемматизации не уступает в простоте и скорости модулю LINGUIST.

Для извлечения МИ, используются маски категорий МИ, приведенные в таблице 4.

Определение значения одной из категорий, например рода, происходит путём применения операции побитового «и» значения МИ и маски категорий. Если словоформе категория не присуща, то результат операции побитового «и» её МИ и маски этой категории будет равен 0. Приведем примеры извлечения значений отдельных категорий морфологической информации.

*Пример 3.* Определение значения категории «род» словоформы:

MI & rod\_mask

Результат: \_Masculine, \_Feminine, \_Neuter или 0.

*Пример 4.* Определение падежа слова, с морфологической информацией MI:

MI & case\_mask

Результат – 0, если слово не склоняется по падежам (глагол, наречие и т.д), или одно из численных значений категории «падеж», приведенных в таблице 3.

Зная МИ слов предложения, можно проводить его синтаксический анализ (СА).

Таблица 3. – Представление МИ в модуле лемматизации

| Значение   |                      | Категория                    | Значение                 |
|------------|----------------------|------------------------------|--------------------------|
| Числовое   | Макроопределения     |                              |                          |
| 0x00000001 | Nominative           | Падеж                        | Именительный             |
| 0x00000002 | Genitive             |                              | Родительный              |
| 0x00000003 | Dative               |                              | Дательный                |
| 0x00000004 | Accusative           |                              | Винительный              |
| 0x00000005 | Ablative             |                              | Творительный             |
| 0x00000006 | Locative             |                              | Предложный               |
| 0x00000008 | Masculine            | Род                          | Мужской                  |
| 0x00000010 | Feminine             |                              | Женский                  |
| 0x00000018 | Neuter               |                              | Средний                  |
| 0x00000020 | Singular             | Число                        | Единственное             |
| 0x00000040 | Plural               |                              | Множественное            |
| 0x00000080 | Past                 | Время, наклонение            | Прошедшее                |
| 0x00000100 | FutureContinuis      |                              | Настоящее/будущее        |
| 0x00000180 | Imperative           |                              | Повелительное наклонение |
| 0x00000200 | FaceFir              | Лицо                         | 1-е                      |
| 0x00000400 | FaceSec              |                              | 2-е                      |
| 0x00000600 | FaceThi              |                              | 3-е                      |
| 0x00000800 | Active               | Залог                        | Действительный           |
| 0x00001000 | Passive              |                              | Страдательный            |
| 0x00002000 | ComparativeFormOfAdj | Степень сравнения, краткость | Сравнительная            |
| 0x00004000 | ExellentFormOfAdj    |                              | Превосходная             |
| 0x00006000 | ShortFormOfAdj       |                              | Краткая форма            |
| 0x00008000 | Verb                 | Часть речи                   | Глагол                   |
| 0x00010000 | Participle           |                              | Причастие                |
| 0x00020000 | Gerund               |                              | Деепричастие             |
| 0x00040000 | Adjective            |                              | Прилагательное           |
| 0x00080000 | Noun                 |                              | Существительное          |
| 0x00100000 | Pronoun              |                              | Местоимение              |
| 0x00200000 | Numeral              |                              | Числительное             |
| 0x00400000 | AdVerb               |                              | Наречие                  |
| 0x00800000 | Preposition          |                              | Предлог                  |
| 0x01000000 | Conjunction          |                              | Союз                     |
| 0x02000000 | Particle             |                              | Частица                  |
| 0x04000000 | Interjection         |                              | Междометие               |
| 0x08000000 | Animate              | Одушевлен-ность              | Одушевленное             |
| 0x10000000 | NotAnimate           |                              | Неодушевленное           |

Таблица 4. – Маски категорий МИ

| Числовое значение | Макроопределение    | Маска категории              |
|-------------------|---------------------|------------------------------|
| 0x00000007        | case_mask           | Падеж                        |
| 0x00000018        | rod_mask            | Род                          |
| 0x00000060        | count_mask          | Число                        |
| 0x00000180        | time_mask           | Время, наклонение            |
| 0x00000600        | face_mask           | Лицо                         |
| 0x00001800        | active_passive_mask | Залог                        |
| 0x00006000        | adjfrm_mask         | Степень сравнения, краткость |
| 0x07FF8000        | part_of_speech_mask | часть речи                   |
| 0x18000000        | animate_mask        | Одушевлённость               |

## Определение подчинительной связи между словами

Задача СА состоит в том, чтобы, используя МИ о словоформах, полученную на этапе МА, и сведения из словаря, построить синтаксическую структуру входного предложения (осуществить разбор предложения). Синтаксическая структура отражает синтаксические связи, существующие между словами в предложении. Так как после этапа МА слово в общем случае представлено в виде совокупности вариантов морфологического анализа, то СА должен устанавливать синтаксические отношения не между словами (или словосочетаниями), а между конкретными вариантами МА (или последовательностями этих вариантов).

Между словами существует 2 вида связи: сочинительная и подчинительная.

**Подчинение**, или **подчинительная связь** — отношение синтаксического неравноправия между словами в словосочетании и предложении, а также между предикативными частями сложного предложения.

В такой связи один из компонентов (слов либо предложений) выступает как главный, другой — как зависимый.

Подчинительной связи в словосочетании и простом предложении противопоставляется **сочинительная связь** - отношение синтаксического (грамматического) равноправия между словами в простом предложении, а также между предикативными частями сложного предложения.

Виды подчинительной связи в словосочетании и предложении:

**Согласование** — вид подчинительной связи, при котором зависимое слово уподобляется в выражении грамматических значений (в роде, числе, падеже) господствующему слову; бывает полное и неполное согласование.

*Примеры:* маленький мальчик, летним вечером; наш врач, на озере Байкал.

**Управление** — вид подчинительной связи, при котором зависимое слово (имя существительное или его эквивалент) ставится в определённой форме косвенного падежа без предлога или с предлогом; главным словом при такой связи может быть слово любой части речи, а зависимым - существительное (включая местоимение-существительное, количественное и собирательное числительное).

*Примеры:* писать стихи, вера в победу, довольный ответом.

**Примыкание** — вид подчинительной связи, при котором подчинённое слово, будучи неизменяемой частью речи или словоформой (инфинитивом или деепричастием, которые синтаксически ведут себя как самостоятельные слова), изолированной от системы падежей, свою зависимость от господствующего слова выражает только местоположением и смыслом (порядком слов, лексически и интонацией). Главным словом при этом может быть глагол, существительное, прилагательное, количественное числительное, а также - при сочетаниях с наречием или компаративом - местоимение-существительное. При этом виде связи формальным показателем зависимости служит сама неизменяемость зависимого слова, а внутренним, семантическим показателем - возникающие отношения. ∴

*Примеры:* внимательно слушать, весьма интересный, слишком поздно, бежать быстро, поворот направо, цвет беж, шинель внакидку, золотистый сбоку, шестой слева, трое наверху, приказ наступать, решиться уехать, поступить умнее, люди постарше, кто-нибудь поопытнее.

В таблице 5 приведены правила, которые будем использовать для определения, есть ли подчинительная связь между двумя словами. В этой таблице характеристики главного слова даны с индексом  $_1$ , а характеристики зависимого слова – с индексом  $_2$ . Если ячейка в таблице 5 не заполнена, значит, характеристика, которую она описывает, не имеет значения при определении связи или не присуща слову. Ячейка, в которой показано равенство некоторой характеристики главного и зависимого слова (например,  $G_1 = G_2$ ), обозначает, что связь возможна только при совпадении значений этой характеристики у главного и зависимого слов. В ряде случаев главное слово требует от зависимого определенной формы. В этом случае, обозначение требуемой формы приводится в соответствующей ячейке.

В ряде случаев подчинительная связь между двумя словами есть только тогда, когда главное слово предшествует подчиненному. В этом случае говорят, что для данной связи обязателен прямой порядок слов. Обязательность прямого порядка слов для каждого из рассматриваемых нами вида подчинительной связи указана в столбце «Прямой порядок слов» таблицы 5. Некоторые виды подчинительной связи между двумя словами могут не зависеть от положения слов в предложении и наличия слов, разделяющих главное и зависимое слово. Например, при связи глагола с наречием, одно из слов может находиться в начале предложения, а второе – в конце, они могут быть разделены другими словами. Если среди этих слов нет других глаголов, можем с уверенностью сказать, что наречие подчинено глаголу. Если между двумя словами возможна синтаксическая связь независимо от того, какими словами они разделены, в столбце «Часть речи допустимых слов-разделителей» таблицы 5 указано значение «Любая». В противном случае указано, при каких словах разделителях связь возможна.

Наличие подчинительной связи между предлогом и существительным будем определять в зависимости от падежа существительного, как показано в таблице 6.

Приведенных данных достаточно для написания алгоритма определения являются ли два слова связанными подчинительной связью. Он состоит в проверке того, является ли первое слово главным и является ли второе слово главным. Если ни одно из слов не является главным, связи между словами нет.

## Грамматики

**Формальная грамматика** или просто **грамматика** в теории формальных языков — способ описания формального языка, то есть выделения некоторого подмножества из множества всех слов некоторого конечного алфавита. Различают порождающие и распознающие (или аналитические) грамматики — первые задают правила, с помощью которых можно построить любое слово языка, а вторые позволяют по данному слову определить, входит оно в язык или нет.

**Терминал** (терминальный символ) — объект, непосредственно присутствующий в словах языка, соответствующего грамматике, и имеющий конкретное, неизменяемое значение (обобщение понятия «буквы»). В формальных языках, используемых на компьютере, в качестве терминалов обычно берут все или часть стандартных символов ASCII — латинские буквы, цифры и спец. символы.

**Нетерминал** (нетерминальный символ) — объект, обозначающий какую-либо сущность языка (например: формула, арифметическое выражение, команда) и не имеющий конкретного символического значения.

**Выводом** называется последовательность строк, состоящих из терминалов и нетерминалов, где первой идет строка, состоящая из одного стартового нетерминала, а каждая последующая строка получена из предыдущей путем замены некоторой подстроки по одному (любому) из правил. Конечной строкой является строка, полностью состоящая из терминалов, и следовательно являющаяся словом языка. Существование вывода для некоторого слова является критерием его принадлежности к языку, определяемому данной грамматикой.

По **иерархии Хомского**, грамматики делятся на 4 типа, каждый последующий является более ограниченным подмножеством предыдущего (но и легче поддающимся анализу):

- тип 0. неограниченные грамматики — возможны любые правила;
- тип 1. контекстно-зависимые грамматики — левая часть может содержать один нетерминал, окруженный «контекстом» (последовательности символов, в том же виде присутствующие в правой части); сам нетерминал заменяется непустой последовательностью символов в правой части;
- тип 2. контекстно-свободные грамматики — левая часть состоит из одного нетерминала;
- тип 3. регулярные грамматики — более простые, эквивалентны конечным автоматам.

Таблица 5 – Правила определения наличия подчинительной связи между двумя словами

| Часть речи       |                           | Род                             |                | Число                           |                | Падеж                           |                | Время          | Лицо           | Часть речи допустимых слов-разделителей | Прямой порядок слов |
|------------------|---------------------------|---------------------------------|----------------|---------------------------------|----------------|---------------------------------|----------------|----------------|----------------|-----------------------------------------|---------------------|
| PS <sub>1</sub>  | PS <sub>2</sub>           | G <sub>1</sub>                  | G <sub>2</sub> | N <sub>1</sub>                  | N <sub>2</sub> | C <sub>1</sub>                  | C <sub>2</sub> | T <sub>1</sub> | F <sub>1</sub> |                                         |                     |
| Глагол           | Наречие                   |                                 |                |                                 |                |                                 |                |                |                | Любая                                   |                     |
|                  | Предлог                   |                                 |                |                                 |                |                                 |                |                |                | Любая                                   | Обязателен          |
|                  | Деепричастие              |                                 |                |                                 |                |                                 |                |                |                | Любая, кроме глагола                    |                     |
|                  | Существительное           | G <sub>1</sub> = G <sub>2</sub> |                | ед.                             | ед.            |                                 | им.            | прош.          |                | Любая                                   |                     |
|                  |                           |                                 |                | мн.                             | мн.            |                                 |                |                |                |                                         |                     |
|                  |                           |                                 |                | N <sub>1</sub> = N <sub>2</sub> |                |                                 |                |                |                |                                         | наст./буд.          |
|                  |                           |                                 |                |                                 |                |                                 | Р, В, Д, Т     |                |                |                                         |                     |
| Существит ельное | Прилагательное, причастие | G <sub>1</sub> = G <sub>2</sub> |                | ед.                             | ед.            | C <sub>1</sub> = C <sub>2</sub> |                |                |                | Прилагательное, причастие               |                     |
|                  | Прилагательное, причастие |                                 |                | мн.                             | мн.            |                                 |                |                |                |                                         |                     |
|                  | Существительное           |                                 |                |                                 |                |                                 | род.           |                |                |                                         | Обязателен          |
| Предлог          | Существительное           |                                 |                |                                 |                |                                 | *              |                |                | Прилагательное, причастие               | Обязателен          |

Условные обозначения: ед. – единственное число; мн. – множественное число; прош. – прошедшее время; наст./буд. – настоящее или будущее время; им. – именительный падеж; род. – родительный падеж; 3-е л. – 3-е лицо; \* – определяется по таблице 6.

Таблица 6 – Предлоги и падежи, с которыми они употребляются

| Падежи      | Предлоги                                                                                                                                                                    |                                  |                  |
|-------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------|------------------|
|             | с одним падежом                                                                                                                                                             | с двумя падежами                 | с тремя падежами |
| Родительный | без, близ, ввиду, возле, вокруг, в отношении, впереди, после, прежде, в продолжение, до, для, вследствие, в течение, из, из-за, кроме, около, от, подле, по мере, против, у | между                            | с                |
| Дательный   | благодаря, вопреки, к, согласно, соответственно                                                                                                                             |                                  | по               |
| Винительный | несмотря на, про, с, сквозь, через (чрез)                                                                                                                                   | в (во), за, на, о, об (обо), под | по, с            |

|              |                                                                    |                    |    |
|--------------|--------------------------------------------------------------------|--------------------|----|
| Творительный | в соответствии с, над (надо), перед (предо), в связи с, согласно с | между, за, под     | с  |
| Предложный   | при                                                                | в, на, о (об, обо) | по |