# Морфологический модуль EP.Morphology

Морфология реализуется в одной сборке EP.Morphology.dll, которую нужно подключить к своему проекту. Все элементы находятся в пространстве имён EP.Morph.

Пример использования в коде:

```
using EP.Morph;

// инициализируем движок нужными языками
Morphology.Initialize(MorphLang.EN | MorphLang.RU | MorphLang.UA);

// запускаем на тексте text
List<MorphToken> tokens = Morphology.Process(text);

// получили токены с морфологическими вариантами
foreach (MorphToken tok in tokens)
       Console.WriteLine(tok.ToString());
```

Исходный текст разбивается на токены, далее морфологический разбор производится для каждого токена отдельно, независимо от окружающих его токенов.  То есть это POS-Tagger (Part of Speech), выдающий для каждого токена-словоформы всевозможные морфологические варианты. Вопрос интерпретации и уменьшения множества вариантов решается вне POS-Tagger. Например, ядро SDK Pullenti при анализе именных групп оставляет только непротиворечивые корневому слову варианты и т.п. Отметим следующее:

- Поддерживаются русский, украинский и английский язык (для белорусского языка пока есть только токенизация);
- Для неизвестных слов выдаются варианты;
- Производится лексикографическая корректировка. Например, если в слове на кириллице одна буква заменена на латинскую, аналогичную по внешнему виду, то производится соответствующая корректировка. Также корректируются буквы с ударениями и т.п., замена апострофа на «ъ» (об`явление), «ё» на «е» и множество экзотических случаев (например, некоторые извращенцы пишут букву Ы через Ь и I);
- Для токена предлагается вариант леммы (нормальная форма);
- Есть функция получения для нормальной формы слова все варианты словоформ;
- При первом запуске движок инициализирует в памяти словари, на что требуется несколько секунд. Чтобы избежать задержку при первом анализе, рекомендуется принудительно вызывать Morphology.Initialize() при старте программы с указанием нужных языков, для которых загружать словари;

- Словарь для каждого языка занимает в памяти от 100 до 150 Мб, словари можно динамически загружать и выгружать через функции LoadLanguages и UnloadLanguages, однако для каждого текста это может быть неэффективно – процедура занимает несколько секунд;

Морфанализ производится через функцию Process статического класса Morphology, которая на выходе выдаёт список токенов List<MorphToken>.

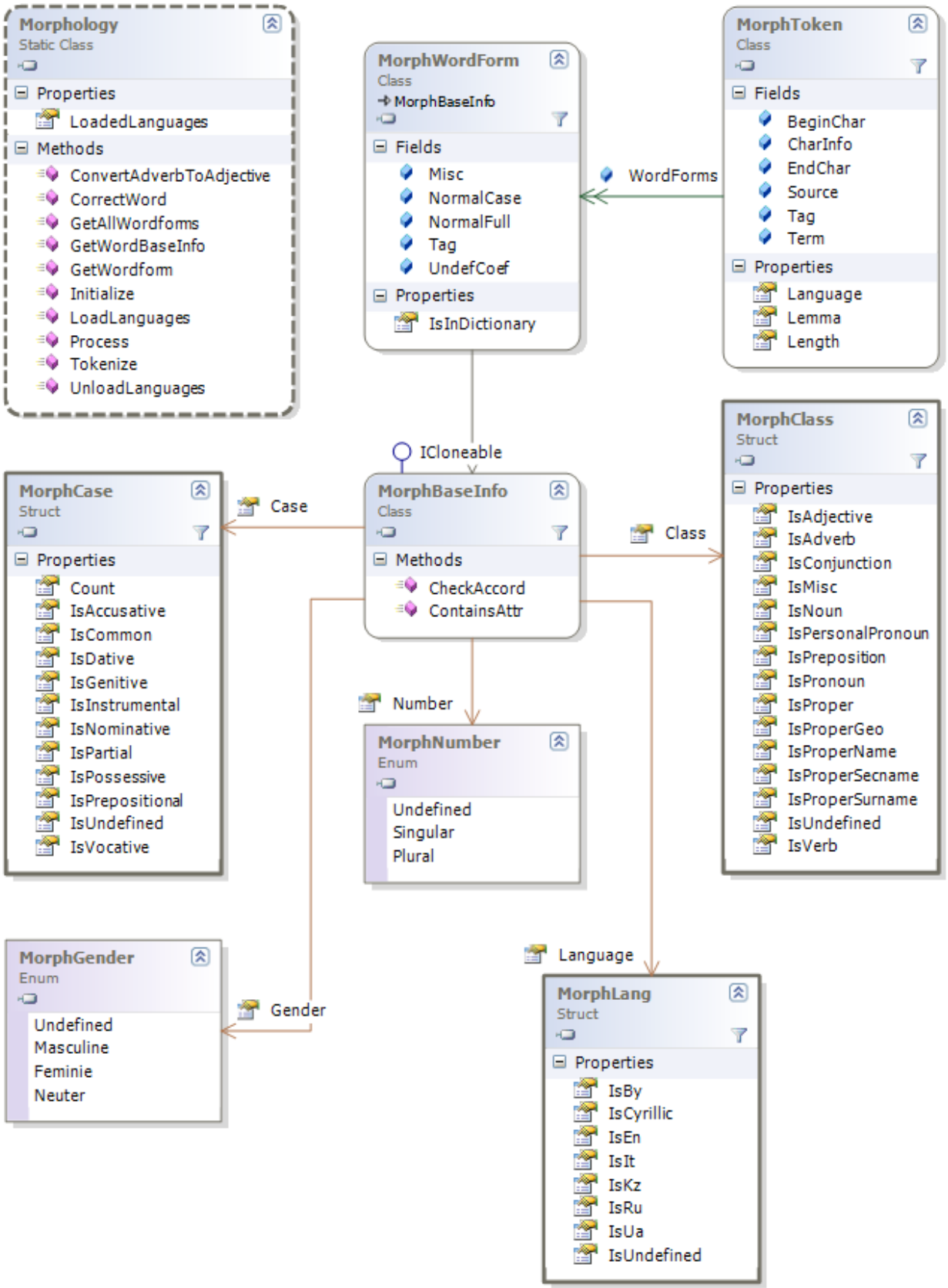
Токен содержит смещение в символах (BeginChar и EndChar), информацию о символах (CharsInfo, см. предыдущий раздел), Term – преобразованный исходный фрагмент (верхний регистр, замена некоторых символов), и список морфологических вариантов WordForms.

Морфологический вариант MorphWordForm имеет тип (Class словоформы: существительное, прилагательное и т.п.), род (Gender), падеж (Case), число (Number) и язык (Language), а также:

- IsInDictionary – признак того, есть ли словоформа в словаре, или этот вариант был создан для неизвестного слова;
- NormalCase – вариант словоформы в именительном падеже;
- NormalFull – вариант в именительном падеже и единственном числе;
- Misc – дополнительная морфологическая информация (см. help);

Токен MorphToken имеет свойство Lemma, которое выбирает из всех словоформ нормальную форму по фиксированному алгоритму. Правил довольно много, например, существительное имеет преимущество перед глаголом, у глаголов обрезается «-СЯ» и т.п. Например, для исходного слова «стекла» с его двумя словоформами (стекло и стечь) леммой будет «СТЕКЛО». В принципе, для поисковых систем такого рода нормализации вполне достаточно. Если предлагаемый алгоритм формирования лемм не устраивает, то можно реализовать свой на основе анализа списка WordForms.

На диаграмме указаны основные классы и структуры, задействованные в морфологии и расположенный в пространстве имён EP.Morph:


Morphology
Static Class
Properties
LoadedLanguages
Methods
ConvertAdverbToAdjective
CorrectWord
GetAllWordforms
GetWordBaseInfo
GetWordform
Initialize
LoadLanguages
Process
Tokenize
UnloadLanguages
MorphWordForm
Class
MorphBaseInfo
Fields
Misc
NormalCase
NormalFull
Tag
UndefCoef
Properties
IsInDictionary
WordForms
MorphToken
Class
Fields
BeginChar
CharInfo
EndChar
Source
Tag
Term
Properties
Language
Lemma
Length
ICloneable
MorphBaseInfo
Class
Methods
CheckAccord
ContainsAttr
Case
Class
Number
MorphCase
Struct
Properties
Count
IsAccusative
IsCommon
IsDative
IsGenitive
IsInstrumental
IsNominative
IsPartial
IsPossessive
IsPrepositional
IsUndefined
IsVocative
MorphClass
Struct
Properties
IsAdjective
IsAdverb
IsConjunction
IsMisc
IsNoun
IsPersonalPronoun
IsPreposition
IsPronoun
IsProper
IsProperGeo
IsProperName
IsProperSecname
IsProperSurname
IsUndefined
IsVerb
MorphNumber
Enum
Undefined
Singular
Plural
MorphGender
Enum
Undefined
Masculine
Feminie
Neuter
Gender
Language
MorphLang
Struct
Properties
IsBy
IsCyrillic
IsEn
IsIt
IsKz
IsRu
IsUa
IsUndefined