Морфология

Морфологията (от гръцки морфе – форма, логос – наука) като дял от лингвистиката е наука, която изучава формите на думите – техния строеж и значение.

За разлика от лексикологията, където думата се възприема като неделима единица, при морфологията думата се изучава като структура, изградена от по-малки значещи сегменти, наречени морфеми.

Като дял от граматиката морфологията е наука, която изучава формата на думите, строежа им, правилата за тяхното образуване и свързаните с формата граматически значения. Лексикалното значение е основно за думата, докато граматическото е допълнително. Напр. думите ходя, ходеше, ще ходя, бих ходил, ходил е, ходиха имат едно общо лексикално значение (действието бягане) и различни граматически значения (род, число, време, вид, залог и наклонение на глагола).

Според най-общата класификация на думите (семантико-граматична, т.е едновременно според лексикалното значение и според формалните признаци) думите се делят на няколко части на речта. Например частите на речта в българския език са :Съществителни имена, Прилагателни имена, Числителни имена, Глаголи, Местоимения, Наречия, Предлози, Съюзи, Междуметия, Частици.

Всяка от тези части има абстрактно значение, което обединява думите в нея и я различава от останалите (например глаголът означава действие или състояние, а прилагателното показва качество на даден предмет или го съотнася към друг). Всяка част притежава определени граматични категории (род, лице, число, вид, спрежение, наклонение и др.) Всяка категория има свои специфични формални признаци – суфикси (определени морфеми), които „носят“ граматическото значение.

Граматическите категории се делят основно на два типа: именни (категории, които притежават имената) и глаголни (категории, които притежават глаголите). Например в българския език съществуват следните граматически категории:

1. Именни категории:

- род (мъжки, женски, среден)
- число (единствено, множествено)
- определеност (определено, неопределено)

- степен (положителна, сравнителна, превъзходна)
- падеж (именителен, винителен, дателен, звателен)

2. Глаголни категории

- лице (първо, второ, трето)
- число (единствено, множествено)
- време (сегашно, бъдеще, бъдеще предварително, минало свършено, минало несвършено, минало неопределено, минало предварително, бъдеще в миналото, бъдеще предварително в миналото)
- вид (свършен, несвършен)
- залог (деятелен, страдателен)
- наклонение (изявително, повелително, условно)
- евиденциалност (индикатив, конклузив, ренаратив (или преизказни форми), дубитатив, адмиратив)

Например в турския език няма разлика в трето лице ед. ч.

Някои форми имат както именни, така и глаголни граматични категории. Например глаголната форма „ходещ“ има не само характерните за глаголите граматични значения (единствено число, сегашно време, несвършен вид, деятелен залог), но и някои характерни за имената (мъжки род, неопределена форма).

В английския език например някои от думите за множествено число получават окончание s/ies, докато други означават едновременно ед. и мн. ч. fish, deer и т.н или mouse/mice, man/men, woman/women или пък използват символи за съкращаване I'm beautiful.

Многозначност на думата

Способността на думите да притежават повече от едно значение се нарича многозначност или полисемия. Многозначните думи съставят около една пета част от лексикалните единици на съвременния български език. Най-голямо е количеството на двузначните думи.

Например многозначната дума *мост* реално съществува в три лексико-семантични варианта: 1/съоръжение, по което се преминава над река, пропаст и т.н.; 2/метална пластинка със зъби по нея; 3/гимнастическо упражнение, при което тялото заема дъгообразно положение.

Омонимите са с еднакъв звуков състав, но различни по значение. Според степента на звуково съвпадение в различните им граматични форми и принадлежността им към една или различни части на речта омонимите се делят на пълни и непълни /частични/. Пълни са тези, които са от една и съща част на речта и всичките им граматични форми съвпадат /например /кран1- а, -ът, -ове/ ‘приспособление, с което се пуска течност или газ от резервоар или тръба’ и /кран2 -а, -ът, -ове/ ‘машина за вдигане и преместване на тежести’/. Непълни са омонимите при които няма съвпадение между всичките им граматични форми /например /дроб1 –а, -ът, -ове/ ‘вътрешен орган у човек или животно/ и дроб2 –та, -и/’число, което е част от единицата’/.

Омографите имат еднакъв звуков състав, но ударението им е върху различни срички /въ’лна – вълна’; па’ра-пара’/. Омографите съвпадат само в писмената реч, тъй като не е прието да се отбелязва мястото на ударението.

Паронимите са близки по звуков състав думи, но различни по значение /афект-ефект, Ботеви-ботевски, група-трупа, превързвам-привързвам, фактура-фрактура/. Реално изявената паронимия има индивидуален характер, понеже зависи от езиковата култура на говорещото лице. Паронимите могат да са от една и съща област в живота /адресат-адресант/ и пароними от различни области /хипотеза-хипотенуза/.

Синонимите са думи с различен звуков състав, но близки по значение /знаме, флаг, трикольор, байрак – те се отличават помежду си по експресивната си окраска и по принадлежността си към различни стилове и жанрове на речта./. Синоними, които са напълно еднакви по значение се наричат абсолютни синоними или лексикални дублети. /виждам, гледам, съзирам/.

Антонимите са с различен звуков състав и са противоположни по значение /болен-здрав, висок-нисък, богат-беден/. Те са думи, които заемат крайни /полярни/ позиции в една семантична парадигма. Антонимични двойки образуват думи само от една и съща част на речта.

Промяната на думата от една словоформа в друга става чрез промяна на структурата на думата, т.е. чрез прибавяне и отнемане на морфеми (суфикси). Морфемата е най-малката значеща единица в езика. Напр.: морфемата -ски е характерна за относителните прилагателни – морски, планински, тя има абстрактното значение на съотнасяне. Морфемата -а е

характерна за съществителните от женски род – *жена, гора*, тя има най-общо женско родово значение. Фонемите например са лишени от семантично съдържание, с изключение на възклицателните „О!“, „А!“ и други, които са думи и присъстват като речникови единици в езика.

Морфемиката е дял от морфологията, чийто обект на изучаване са конкретно морфемите. Морфемният анализ означава разделянето на думата на морфеми (представка, корен, наставка, окончание, определителен член). Напр.: думата предизвикателството е съставена от представките пред-из-, корена -вик-, наставките -а-тел-, окончанието -ство- и определителния член -то. Не бива да се бъркат със сричките, които са вид фонетично делене на думата (пре-ди-зви-ка-тел-ство-то).

Ако една от тези морфеми се замени или премахне (пред-из-вик-а-тел-ства) се променя граматическото значение на думата (от среден род, единствено число, определено, в множествено число, неопределено). Изключение е коренът, който рядко променя формата си, тъй като е свързан с лексикалното значение на думата.

Гранична област между Морфологията и Лексикологията е словообразуването. Както показва самият термин то се занимава с начините за образуване на нови думи чрез добавяне на морфеми с лексикално значение. Например чрез заменяне на морфемата из- с морфемата от- в думата изпия се получава новата дума отпия.

Формите *отпия* и *изпия* имат едни и същи граматически значения, но се различават по лексикалното си значение. Следователно те са форми на две различни думи, а не на една и съща.

Богатството от езиково разнообразие прави задачата за разпознаване на текстови формати още по-трудна и изисква комплексен подход за решаването ѝ.

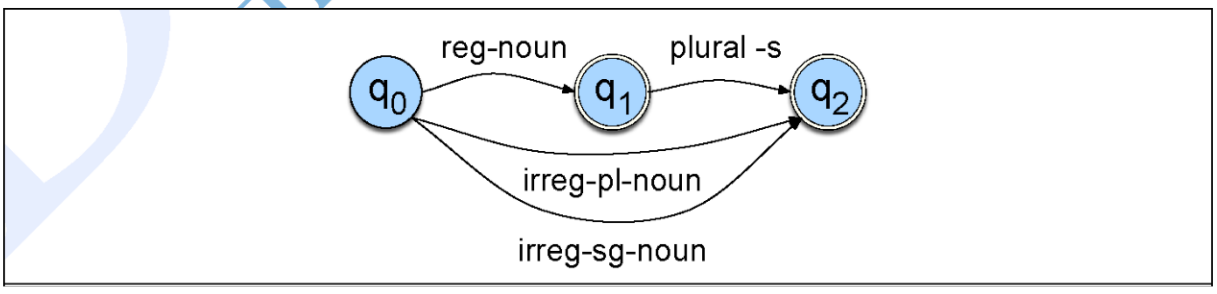
На следващата фигура са представени някои морфологични изменения на английски и испански език.

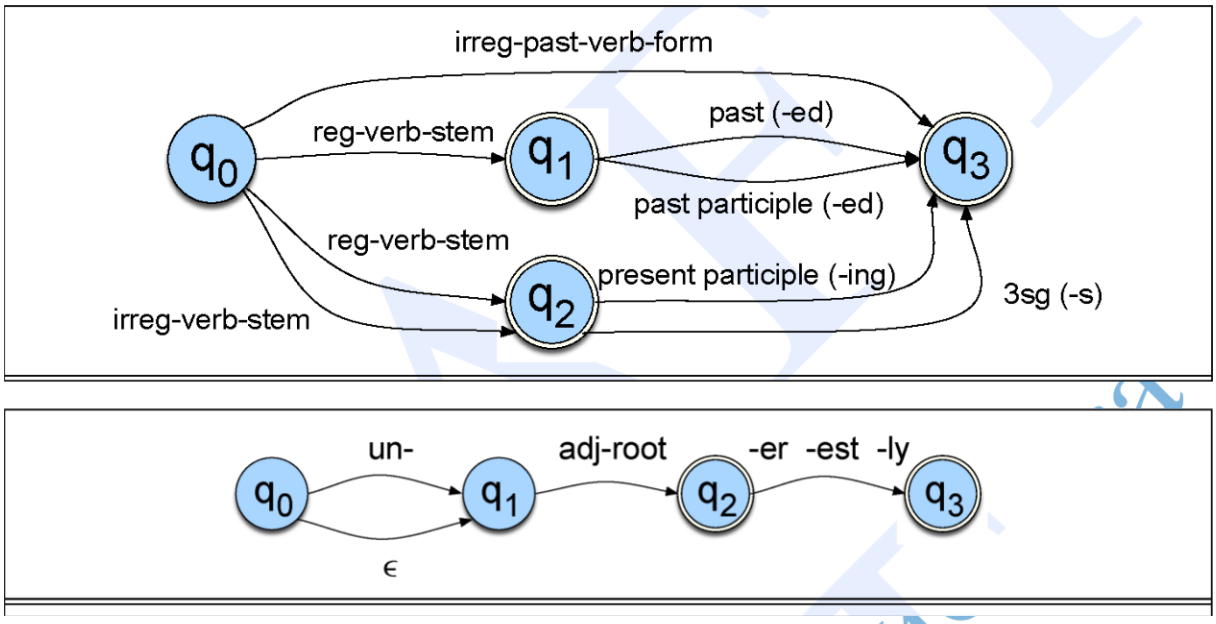
English		Spanish		
Input	Morphologically Parsed Output	Input	Morphologically Parsed Output	Gloss
cats	cat +N +PL	pavos	pavo +N +Masc +Pl	'ducks'
cat	cat +N +SG	pavo	pavo +N +Masc +Sg	'duck'
cities	city +N +Pl	bebo	beber +V +PInd +1P +Sg	'I drink'
geese	goose +N +Pl	canto	cantar +V +PInd +1P +Sg	'I sing'
goose	goose +N +Sg	canto	canto +N +Masc +Sg	'song'
goose	goose +V	puse	poner +V +Perf +1P +Sg	'I was able'
gooses	goose +V +1P +Sg	vino	venir +V +Perf +3P +Sg	'he/she came'
merging	merge +V +PresPart	vino	vino +N +Masc +Sg	'wine'
caught	catch +V +PastPart	lugar	lugar +N +Masc +Sg	'place'
caught	catch +V +Past			

Втората колона съдържа основата на всяка дума, както и различни логически морфо-характеристики. Тези характеристики уточняват допълнителна информация за корена. Испанският има някои функции, които не се срещат в английския; например съществителните lugar и pavo са отбелязани с +Masc (мъжки род). Тъй като испанските съществителни съвпадат по род с прилагателни, познаването на рода на съществителното ще бъде важно за маркиране и синтактичен анализ.

Обърнете внимание, че някои от формите за въвеждане (като caught, goose, canto или vino) ще бъдат двусмислени между различните морфологични анализи.

На следващата фигура е представен автомат с краен брой състояния за думи с корен, производните им или такива с представки и надставки.





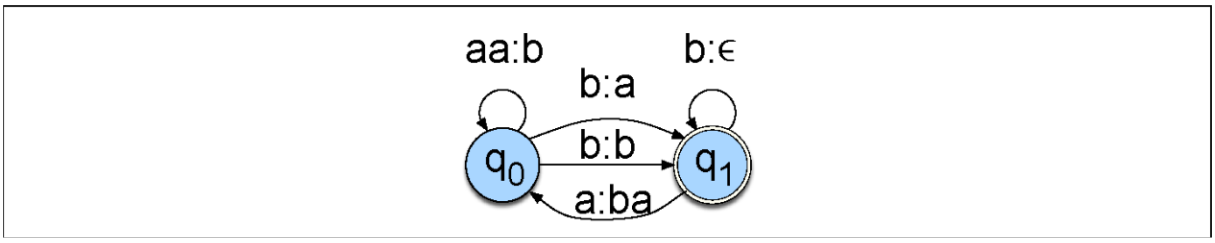
Докато този автомат разпознава всички прилагателни в таблицата по-горе, той ще разпознае и неграматични форми. Трябва настройка за класове от корени и да се посочат възможните им суфикси. Така adj-root1 ще включва прилагателни, които могат да се срещат с un- и -ly (), докато adj-root2 ще включва прилагателни, които не могат и т.н.

Това дава представа за сложността, която може да се очаква само от думи с английски произход без да се отчитат ment, ful,less ness, able и още и още.

Дотук е видно, че FSA могат да представляват морфологична структура на лексикона и могат да се използват. В следващия раздел е представен преобразувателя с ограничено състояние и ще покаже как преобразувателите могат да бъдат приложени към морфологичен анализ.

Макар и с множество настройки, автомата с крайни състояния може да бъде използван за разпознаване на думи. Посредством някои изменения е възможно да се създаде вид краен автомат, който преобразува между два набора от символи или т. нар преобразуватели finite-state transducer (FST).

Интуитивно може да се направи , като се маркира всяка дъга в крайната машина с два символни низа, по един от всяка лента, където всяка дъга е обозначена с входен и изходен низ, разделени с двоеточие.



Следователно FST има по-обща функция от FSA; където FSA дефинира формален език чрез дефиниране на набор от низове, FST дефинира връзка между набори от низове. Друг начин за разглеждане на FST е като на машина, която чете един низ и генерира друг. Ето обобщение на този четиристранен начин на мислене за преобразувателите:

- FST като разпознавател – преобразувател - приема чифт низове като вход и извежда набор, ако двойката низове на езика на двойката низове, и отхвърля, ако не е.
- FST като генератор: машина, която извежда двойки низове на съответния език. Така изходът е потвърден или отхвърлен за двойка изходни низове.
- FST като транслатор: машина, която чете низ и извежда друг низ
- FST като set relater: машина, която изчислява връзките между множествата.

Всички те имат приложения в обработката на реч и език. За морфологичен анализ е представен преобразувател използван за превод, като се вземе входен низ от букви и произвежда изходен низ от морфеми.

Формална дефиниция. FST може да бъде дефиниран със 7 параметъра:
Q краен набор от N състояния $q0, q1, qN-1$
S крайно множество, съответстващо на входната азбука

A краен набор, съответстващ на изходната азбука

$q^o \subset Q$ началното състояние

$F \subset Q$ множеството от крайни състояния

$\delta(q, w)$ преходната функция или преходната матрица между състоянията;

Като се има предвид състояние $q \subset Q$ и низ $w \subset S^*, \delta(q, w)$ връща набор от нови състояния $Q' \subset Q$. Следователно δ е функция от $Q\ x\ S^*$ до 2^Q (защото

има 2^q възможни подмножества на Q). δ връща набор от състояния, а не едно състояние, тъй като даден вход може да е двусмислен в кое състояние се преобразува.

$\sigma(q,w)$ изходната функция, даваща набора от възможни изходни низове за всяко състояние и вход. При дадено състояние $q \subset Q$ и низ $w \subset S^*$, $\sigma(q,w)$ дава набор от изходни низове, всеки низ $q \subset A^*$следователно σ е функция от $Q\ x\ S^*$ до 2^σ

Там където FSA са изоморфни на регулярните езици, FST са изоморфни на регулярните отношения. Регулярните отношения са набори от двойки низове, естествено разширение на регулярните езици, които са набори от низове. Подобно на FSA FST и регулярните отношения са затворени при обединение, въпреки че като цяло не са затворени при разлика, допълване и пресичане (въпреки че някои полезни подкласове на FST са затворени при тези операции; като цяло FST, които не са разширени е по-вероятно да имат такива свойства на затваряне). Освен обединение, FST имат две допълнителни свойства на затваряне, които се оказват изключително полезни

- Инверсия: Инверсията на преобразувател $T\ (T^{-1})$ просто превключва входните и изходните етикети. По този начин, ако T преобразува от входната азбука I към изходната азбука O, T^{-1} преобразува от O в I.
- Композиция: Ако T^{-1} е преобразувател от $I1$ към $O1$ и $T2$ е преобразувател от $O1$ към $O2$, тогава $T1$ или $T2$ се преобразува от $I1$ към $O2$

Инверсията е полезна, защото улеснява конвертирането на FST-as-parser в FST-as-generator.

Композицията е полезна, защото позволява взимане на два преобразувателя, които работят последователно, и заменят с един по-сложен преобразувател. Композицията работи като в алгебрата; прилагането на $T1$ или $T2$ към входна последователност S е идентично с прилагането на $T1$ към S и след това $T2$ към резултата; по този начин

$T1 * T2(S) = T2(T1(S))$. Както е посочено в примера