



Természetesnyelv-feldolgozás

Bevezetés

- **Nyelv- és beszédtechnológia:**
 - írott és a hangzó nyelv feldolgozása
 - nyelvi produktumok előállítása
- **Natural language processing (természetesnyelv-feldolgozás)**
- **Computational linguistics (számítógépes nyelvészet)**
- **Human language technology (nyelvtechnológia)**



Célok

- az ember-ember, az ember-gép kommunikáció hatékonyabbá tétele
- az emberi munkavégzés megkönnyítése újszerű, számítógépes termékek és szolgáltatások biztosításával
- hátrányos helyzetű csoportok (siketek, gyengénlátók, baleset következtében beszédképességüket elvesztők, idegen nyelveket nem tudók) életminőségének javítása





Nyelvtechnológia a mindennapokban

- Helyesírás-ellenőrző
- Webes keresés (Google)
- Fordítóoldalak (Google Translate, webfordítás)
- Címkézés [origo]
- Hanghívás
- Telefonos tudakozó
- ...



Részterületek

- **Nyelvi szintek (elemzés):**
 - szegmentálás
 - morfológia
 - szintaxis
 - szemantika
- **Alkalmazások (pl.):**
 - információ-visszakeresés és –kinyerés
 - gépi fordítás

Mi kell a sikeres elemzéshez?

- Megfelelő program, ill. algoritmus 😊
- Ezek tanító, illetve tesztadatbázist igényelnek -> kézzel annotált korpusz
- Kiértékelés: emberi teljesítményhez mérik



Korpusz

- speciális célokra létrehozott, (gyakran tematikus) adatbázis – „szöveggyűjtemény”
- nyelvészek (szakértők) kézzel bejelölik (annotálják) a szövegekben a releváns információkat
- lehet annotáció nélküli is (statisztika, gyakorisági adatok mérése)
- Magyar Nemzeti Szövegtár
- Szeged Korpusz és Treebank
- Webkorpusz



Mondatszegmentálási problémák

- mondat az a szövegegység, amely nagybetűvel kezdődik és ponttal, kérdőjellel vagy felkiáltójellel zárul
- nem szokványos írásjel:
 - *“Mondottam, ember: küzdj és bízva bízzál!”*
- szöveg tagolása, formázása HTML-ben:
 - *<p>A sofőr közvetített</p><p>A román ügyészség szerint Vantu a csőd előtt jelentős összegeket vont ki az FNI-ből, majd eladta a céget.*
- kisbetűs mondatkezdet:
 - *iPhone-t kért karácsonyra.*
- írásjel a mondat közepén:
 - *A Yahoo! a Yahoo! Inc. amerikai cég védjegye, amellyel egy internetes portált és katalógust üzemeltet.*



A szövegszavakra bontás problémás esetei

- **szövegszó:** a szövegnek az a részlete, amely két szóköz között helyezkedik el, leszámítva az írásjeleket
 - mondatok vagy bekezdések első szava
 - többszavas kifejezések (*kútba esés, Magyar Nemzeti Bank, 3 000 000*)
 - számmal írt sorszámnevek (*1992. évi LVIII. törvény*)
 - pontra végződő rövidítések (*stb.*)
 - írásjelet tartalmazó szavak (*labdarúgó-világbajnokság, McDonald's*)





Morfológiai elemző

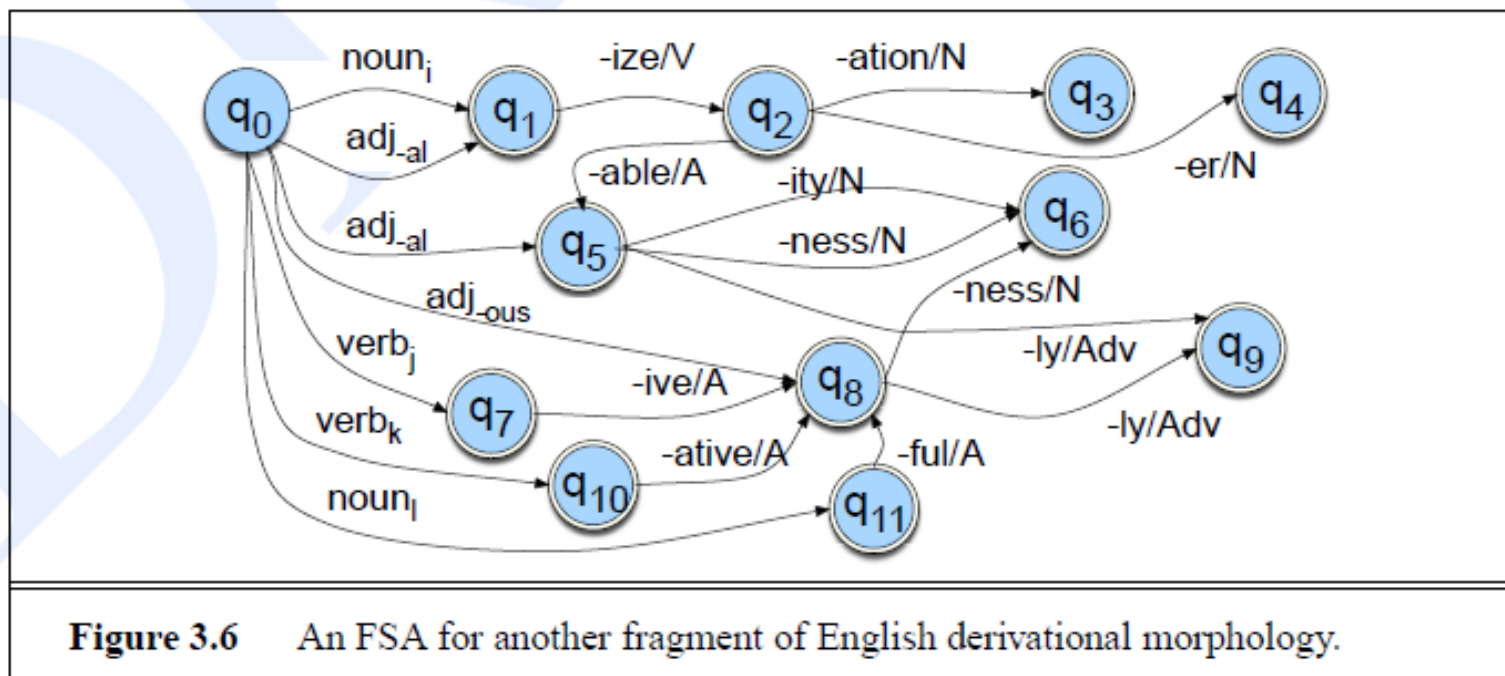
- **Feladata: minden szövegszóhoz hozzárendelni az összes lehetséges elemzését és a hozzájuk tartozó szótöveket**
- **magyar vs. angol**
- **nem gazdaságos szótárban tárolni a szóalakokat**
- **új szavak születnek: összetétel, képzés...**
- **a szótárban való keresés nem elégséges: más módszerek kellenek**



Lemmatizálás

- Szótövesítés: a morfológiai elemző feladata – a szóalakot töre (lemmára) és toldalékokra bontja
- „normális” szavak esetén az elemző szótárra + toldaléklistára hagyatkozik
- ismeretlen (új, szótárban nem szereplő) szavak (pl. a tulajdonnevek) nincsenek benne a szótárban -> csak a toldaléklista elérhető -> lehetséges toldalékok levágása a szó végéről (*Pannon* – *Pann?*)

Automaták a morfológiában



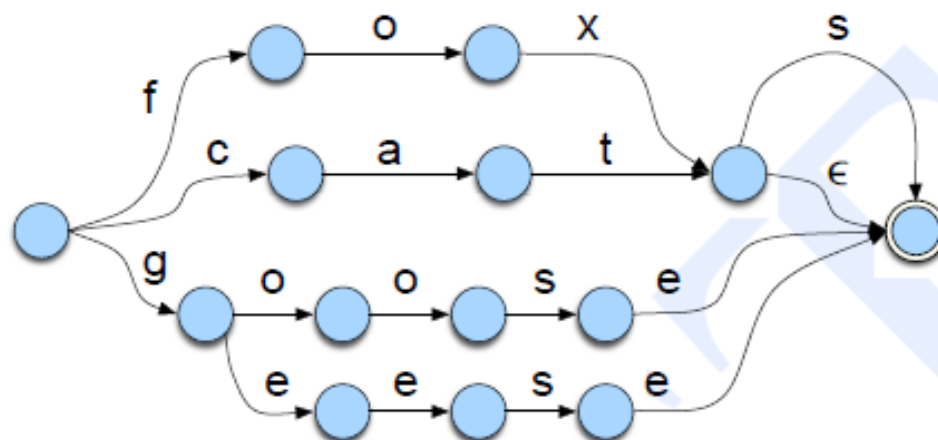


Figure 3.7 Expanded FSA for a few English nouns with their inflection. Note that this automaton will incorrectly accept the input *foxs*. We will see beginning on page 19 how to correctly deal with the inserted *e* in *foxes*.

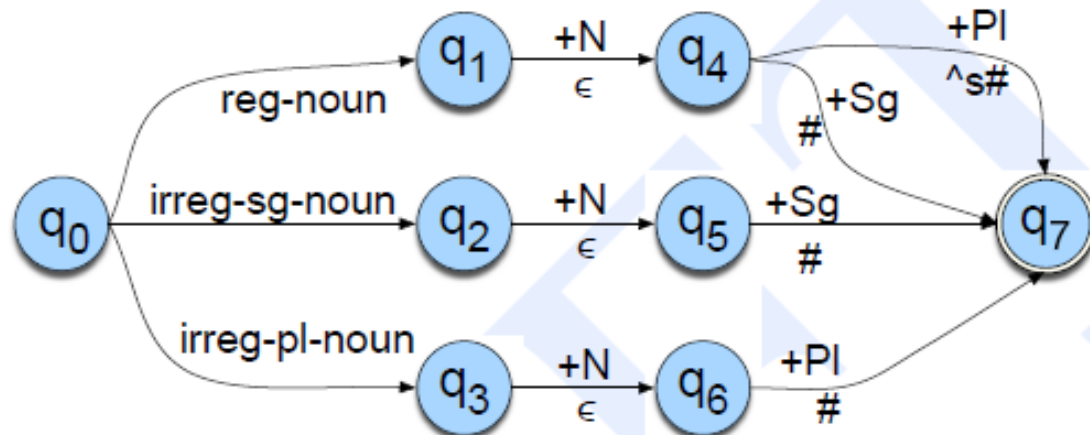


Figure 3.13 A schematic transducer for English nominal number inflection T_{num} . The symbols above each arc represent elements of the morphological parse in the lexical tape; the symbols below each arc represent the surface tape (or the intermediate tape, to be described later), using the morpheme-boundary symbol $\wedge$ and word-boundary marker $\#$. The labels on the arcs leaving q_0 are schematic, and need to be expanded by individual words in the lexicon.

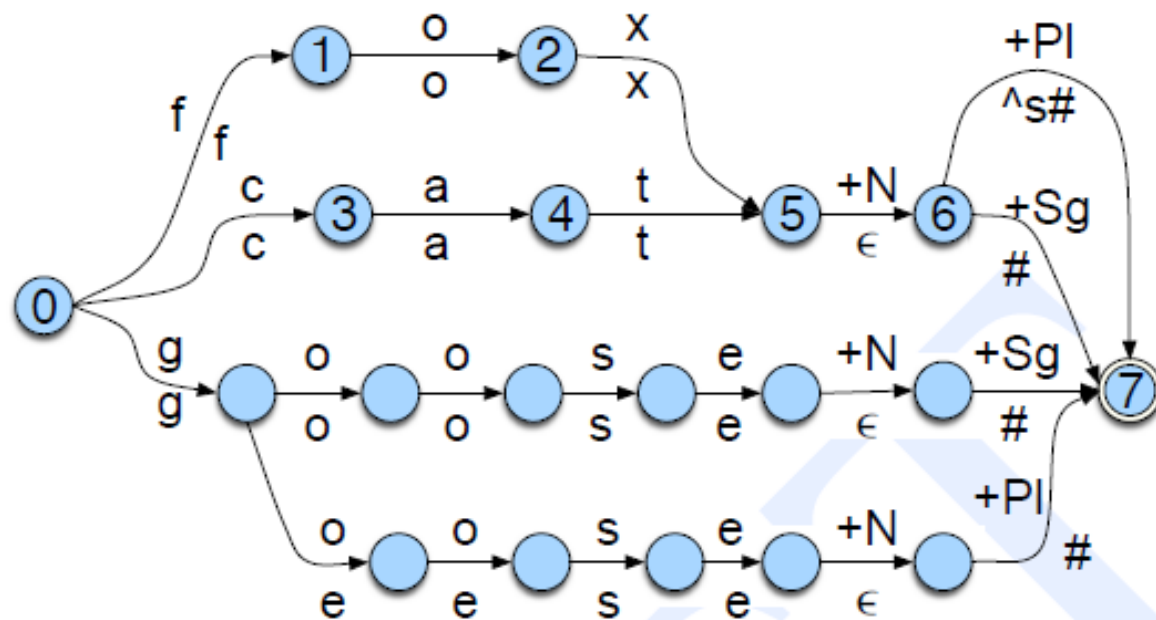


Figure 3.14 A fleshed-out English nominal inflection FST T_{lex} , expanded from T_{num} by replacing the three arcs with individual word stems (only a few sample word stems are shown).

Szófaji egyértelműsítés

- Part-of-speech (POS)-tagging – POS-tagger
- A morfológiai elemző által adott kódok közül kiválasztja az adott környezetbe illőt:

Megcsípett, a szemét!

V DET N-NOM

Hol csípte meg?

ADV V PREV

A szemét.

DET N-ACC

(ADV: határozószó, DET: névelő, N-ACC: tárgyesetű főnév, N-NOM: alanyesetű főnév, PREV: igekötő, V: ige)

- Szekvenciajelölés
- MEMM vagy CRF



Magyar morfológiai elemzők és POS-taggererek

- Hunmorph
- Hunpos
- Humor
- magyarlanc

<http://mokk.bme.hu/resources/hunmorph>

<http://www.inf.u-szeged.hu/rgai/magyarlanc>



Szintaxis

- Szavak között levő kapcsolatok azonosítása gépi úton: adott mondathoz az elemzési fákat megtalálni és kiválasztani a legjobbat
- Parsing – parser
- Sekély vs. mély elemzés
- Részleges elemzés (chunking) – főnévi csoportok (NP-k)





Szintaxiselméletek

- **függőségi nyelvtanok**
- **frázisstruktúra-nyelvtanok**
- **szabad szórendű nyelvek leírására különösen a függőségi nyelvtanok hasznosak**
- **kötött szórendű nyelvekre jók a frázisstruktúra-nyelvtanok**

Szintaktikai többértelműség

- bővítmények csatolása:

Láttam a lányt a távcsővel.

– Kinél van a távcső?

- mellérendelés:

(Sikoltozó (gyerekek és anyukák)) szaladgáltak a játszótéren.

((Sikoltozó gyerekek) és (anyukák)) szaladgáltak a játszótéren.

- szintaktikai egyértelműsítés: az összes lehetséges fa előállítása a mondathoz és a legjobb kiválasztása

Show me the meal on Flight UA 386 from San Francisco to Denver. – 14 elemzési fa





Magyar treebankek és elemzők

- **Szeged Treebank:**
 - frázisstruktúra
 - dependencia
- **Hunpars**
- **MetaMorpho**
- **magyarlanc**

Szemantika

- Tartalmi elemzés automatikusan?
- Szöveg jelentésének megragadása távoli cél...
- Információkinyerést és gépi fordítást nagyban segítően
- Szavak jelentésének azonosítása könnyebb → jelentés-egyértelműsítés (word sense disambiguation, WSD)



Alkalmazások

- Az előző technológiák, algoritmusok, elemzők stb. gyakorlati hasznosítása
- „A jéghegy csúcsa”
- Hétköznapi ember számára is hasznos
- Információ-visszakeresés
- Információkinyerés
- Gépi fordítás





Információ-visszakeresés

- Information retrieval (IR)
- Olyan dokumentumok összegyűjtése, melyek relevánsak az adott keresés szempontjából
- Keresőprogramok
- Szemantikus web
- Kell(ene) hozzá:
 - WSD (*póló*)
 - Lemmatizálás (*foci*, *focinak*, *focival* stb.)
 - Tagadás és bizonytalanság felismerése



Információkinyerés

- Information extraction (IE)
- Nyers strukturálatlan szövegből strukturált információ
- Pl. gazdasági hírekből (teljes mondatokból) rendezett n-esek kinyerése:
 - Eladó-vevő-áru-érték
- Hatékonyabb, de nagyobb szakértői tudást igényel
- Általában doménspecifikus megoldások (üzleti hírek vs. orvosi szövegek)



IE-alkalmazások

- **Névelem-felismerés**
- **Anonimizálás**
- **Orvosi, biológiai dokumentumokból információ kigyűjtése**
- **Véleménykinyerés**
- **Címkézés, kulcsszókinyerés**



NE-felismerés

- **NE = Named Entity ~ tulajdonnevek és azonosítók**
- **Rendszámok, telefonszámok, webcímek... - reguláris kifejezésekkel felismerhetők**
- **Alkalmazásfüggő, mi számít NE-nek (vegyületek, génnevek, páciens neve, orvos neve...)**
- **4 klasszikus kategória: PER (person – személy), ORG (organization – szervezet), LOC (location – hely), MISC (miscellaneous – egyéb)**

Miért kell a NER?

- NE-k különleges bánásmódot igényelnek
- George Bush – György Bokor ?
- Barack Obama – Peach Obama ?
- Kovács János – János Kovács
- ENSZ – UN
- Beijing – Peking
- Anonimizálás: azonos típusúra lecserélni



Orvosi-biológiai IE

- **automatikus BNO-kódolás (betegségek nemzetközi osztályozása)**
- **betegek dohányzási státuszának, káros szenvedélyeinek megállapítása kórlapok alapján**
- **elhízottság és kapcsolódó betegségek megállapítása**
- **kórlapok automatikus osztályozása pl. a fentiek alapján**





Kivonatolás

- **Címkézés / kulcsszókinyerés:** dokumentumokhoz azok tartalmát jól tükröző szavak, kifejezések hozzárendelése [origo]
- **Véleménykinyerés:** fórumokból, blogokból a szerzők véleményének megállapítása egy adott témával kapcsolatban -> piackutatás, fogyasztói visszajelzések megismerése

Gépi fordítás

- **Nehéz feladat: jelentést kell visszaadni, ugyanakkor a másik nyelv morfológiai és szintaktikai szabályainak is meg kell felelni**
- **Jelenleg a minőség (jelentősen) elmarad az emberi fordítástól**
- **DE: bizonyos részterületeken elég**
 - Receptek
 - Időjárás-jelentés
- **Óriási igény lenne rá, főleg Magyarországon ☹**



Domének

- **EU: nyersfordítás készítése, amit emberi erővel ellenőriznek / javítanak**
- **Szépirodalom: nagyon nehéz, megoldhatatlan ☹**
- **Szakszövegek, használati utasítások: elég jól automatizálható, minimális emberi erőforrás**





Módszerek

- Példaalapú: adatbázisban tárolt mondatok/kifejezések közül melyikre hasonlít legjobban
- Szótáralapú: gyenge, rokon nyelvekre lehet jó
- (szintaktikai) transzfer alapú: legtöbb módszer ezt használja, tűrhető megoldások
- Interlingua: köztes nyelvre fordít – „utópia”

Fejlesztések

- MetaMorpho (www.webforditas.hu)
- Hunglish nyersfordító
- Emberi fordítást segítő eszközök fejlesztése:
 - intelligens szótárak
 - lexikai adatbázisok
 - fordítómemóriák
 - párhuzamos korpuszok



Vicces példák

- 2006-ból...
- Összetett szavak:

bányászszív	[miner sucks]
bulvárszíndarab	[boulevard colour][piece]
gyertyamártás	candle sauce
habképző	[foam][derivational suffix]
hajsütés	hair baking
halálnem	death gender
halmajonéz	[dying mayonnaise]
hóhányás	[snow][vomit]
hóhullás	[snow][corpse]
hőképzés	[heat][training]
hőkiütés	[heat][knockout]
időjóslat	[time][prophecy]
+ light verb construction	fényigeszerkezet (2012)



Magyar-angol fordítás

- Nyelvtani hiányosságok:

Ma te **vigyáztál** a gyerekeire, **nem**?

[Today][you] **vigyáztál**[his children],[**gender**]? (MetaMorpho)

Yet today **your are take care** the about your child , **not**? (InterTran)

You **looked after** her children today, **didn't you**? (referencia)



Angol-magyar fordítás

- Többjelentésű szó nem megfelelő jelentésben:

Her children go to the same school as **mine**.

A gyerekei **bányaként** járnak ugyanabba az iskolába. (MetaMorpho)

Gyermekei megy ugyanaz iskola mint **bánya**. (InterTran)

A gyerekei ugyanabba az iskolába járnak, mint az **enyémek**. (referencia)



Angol-magyar fordítás

- Többjelentésű szó nem megfelelő jelentésben:

There are no **biscuits left!**

Kekszek nincsenek **balra!**
(MetaMorpho)

Nincs **kétszersültek bal!** (InterTran)

Nincs **több keksz!** (referencia)

