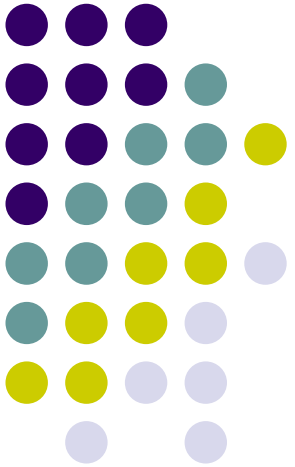
Building a Large Scale LFG Grammar for Turkish

Özlem Çetinoğlu
Sabancı University
İstanbul, Turkey

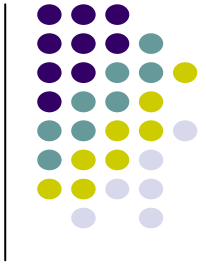
DCU November 2008



Motivation

- Why do we need grammars?
  - to understand and to represent the language in a formal way
  - as a resource
    - machine translation
    - summarization, paraphrasing
    - applications
    - ...

Purpose

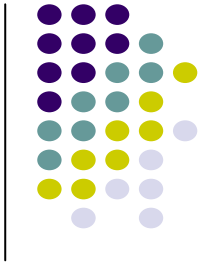


- A large scale grammar for Turkish in LFG formalism
  - using *segments of words* as the building units of rules to explain the linguistic phenomena in a more formal and accurate way
  - paying attention to coverage
  - without leaving aside the interesting linguistic problems to be solved

Turkish LFG Project

- supported by Tübitak (Turkish NSF), 10/2005 – 9/2008
- member of Parallel Grammars (ParGram) Project
  - English, German, French, Japanese, Norwegian
  - Chinese, Urdu, Malagasy, Arabic, Welsh, Hungarian, Tigrinya, Georgian

Outline



- Turkish in General
- Inflectional Groups
- Framework
- Work Accomplished
- Ongoing/Future Work
- Conclusion

Turkish - Morphology

- Agglutinative morphology
- Very productive inflectional and derivational processes

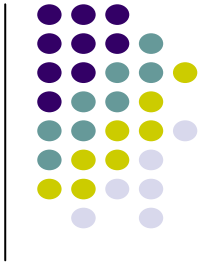
```
ev                  +im   +de   +ki

ev+Noun+A3sg   +P1sg +Loc  ^DB+Adj+Rel
```

‘in my house’

Finite state implementation (Oflazer 1994)

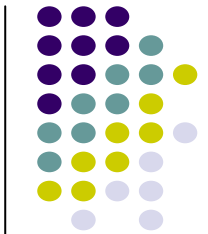
Turkish - Morphology



- In a typical running Turkish text
  - There is an average of 3-4 morphemes per word
  - With an average of 1 derivations per word when high-frequency function words are not considered (Eryiğit and Oflazer 2006)

- Derivational processes play an important role in sentence structure

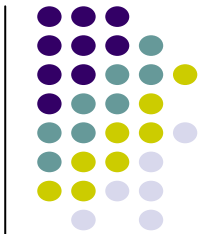
Turkish - Syntax



- Free constituent order in sentence level
  - generally SOV
  - almost no constraints

- The case of a noun phrase determines its grammatical function in the sentence

Representing Morphological Information

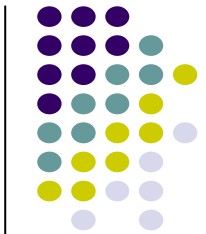


- Each morphological analysis of a word can be represented as a sequence of Inflectional Groups **(IGs)**

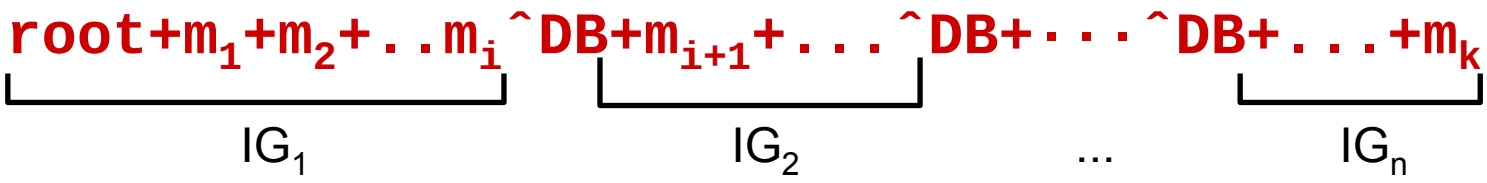
$$\underbrace{root+m_1+m_2+..m_i}_{IG_1}\ \hat{}DB\ \underbrace{+m_{i+1}+...}_{IG_2}\ \hat{}DB+\cdots\hat{}DB\ \underbrace{+...+m_k}_{IG_n}$$

- Each IG_i corresponds to a sequence of inflectional features

Representing Morphological Information

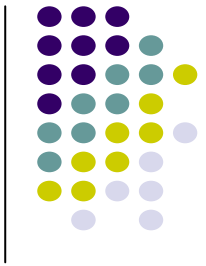


- Each morphological analysis of a word can be represented as a sequence of Inflectional Groups **(IGs)**



- ^DB indicates a derivation boundary
- An IG is typically larger than a morpheme but smaller than a word

Representing Morphological Information



- *canlısı* (the lively one of)

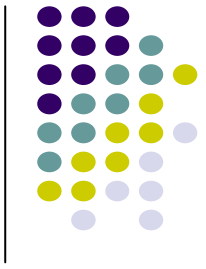
Morphological Analysis:

```
can+Noun+A3sg+Pnon+Nom^DB+Adj+With
                    ^DB+Noun+Zero+A3sg+P3sg+Nom
```

IGs:

1. can+Noun+A3sg+Pnon+Nom
2. +Adj+With
3. +Noun+Zero+A3sg+P3sg+Nom

Inflectional Groups and Syntactic Relations

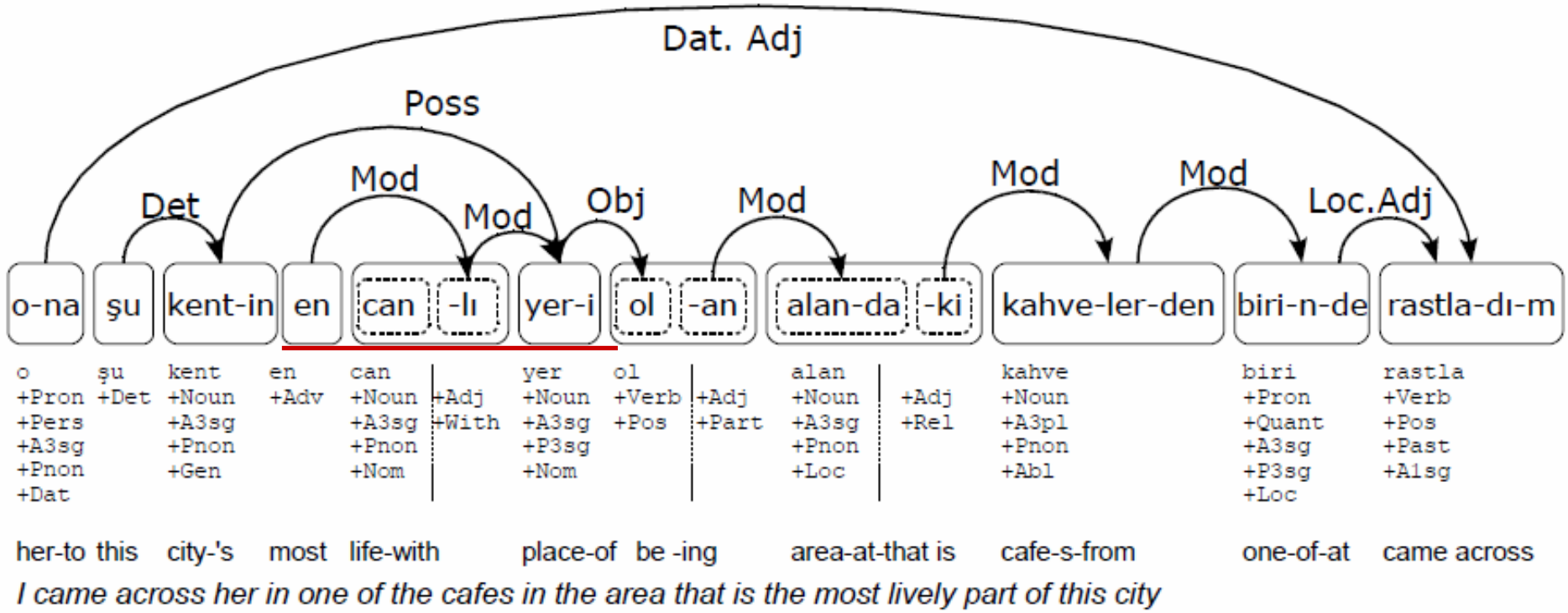


- Why use IGs?

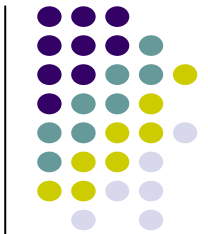
- Syntactic relations are between inflectional groups (IGs), not between words

Inflectional Groups and Syntactic Relations

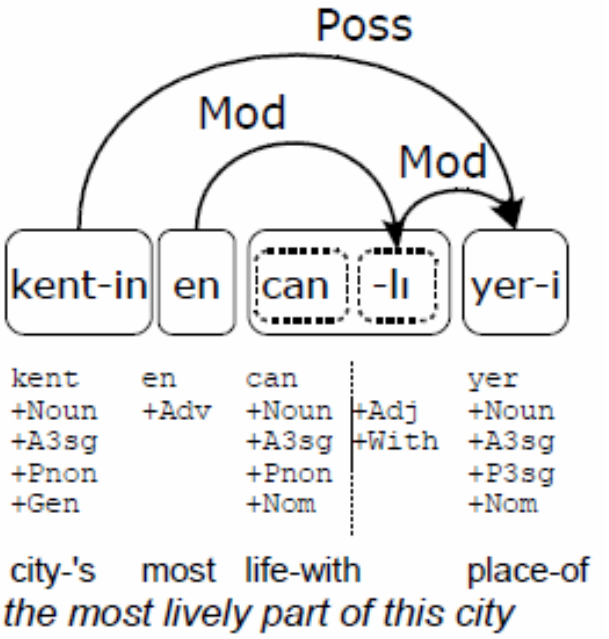
- Heads are almost always to the right



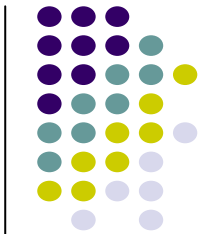
Inflectional Groups and Syntactic Relations



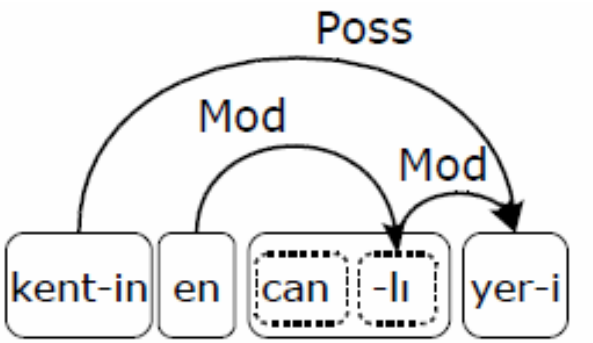
- Adverbial *en* modifies the derived adjective *canlı*
- AP *en canlı* modifies *yeri*
- possessive noun *kentin* modifies *yeri*

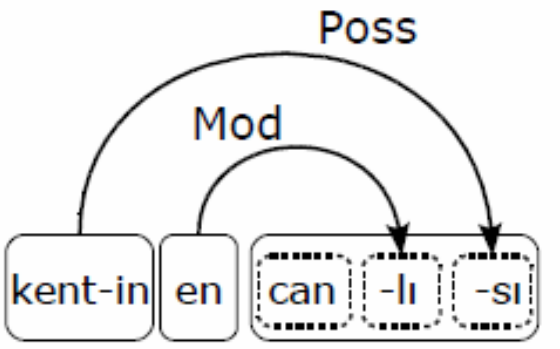


Inflectional Groups and Syntactic Relations

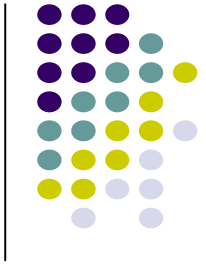


- Adverbial *en* modifies the derived adjective *canlı*
- The modified adjective is derived into a noun
- *kentin* (modifying *yeri* in the first example) modifies derived noun *canlısı*



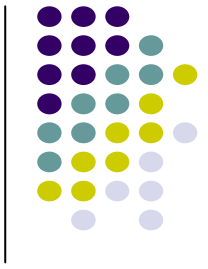


Outline



- Turkish in General
- Inflectional Groups
- Framework
- Work Accomplished
- Ongoing/Future Work
- Conclusion

Framework



- Lexical Functional Grammar (Dalrymple 2001)
  - unification based grammar
  - developed by Kaplan&Bresnan in 1980s

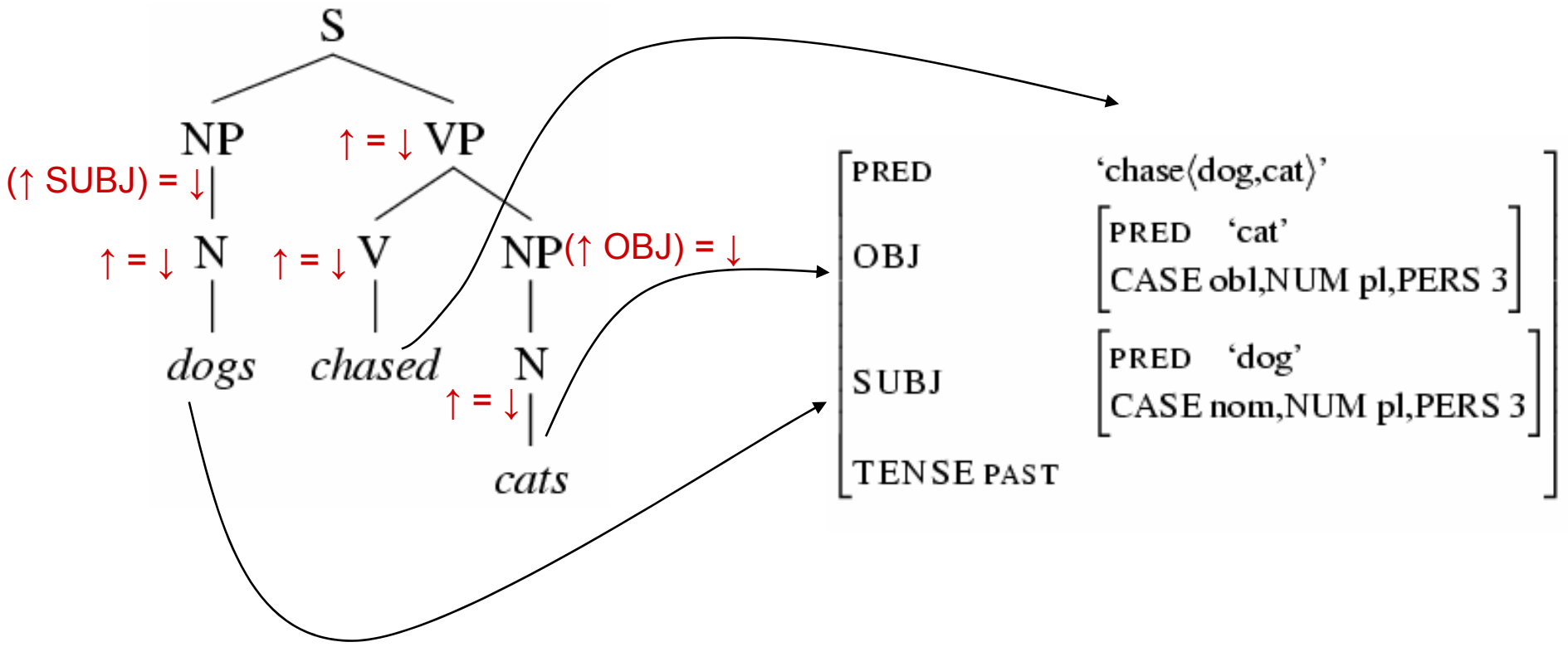
- XLE – Xerox Linguistic Environment (Maxwell and Kaplan 1996)
  - for building LFG grammars
  - efficient, has rich GUI
  - developed at Xerox PARC in 1990s

Lexical Functional Grammar

- Representing syntax in two levels
- Constituent Structure
  - Context free phrase structure trees
  - Order and grouping → Language specific

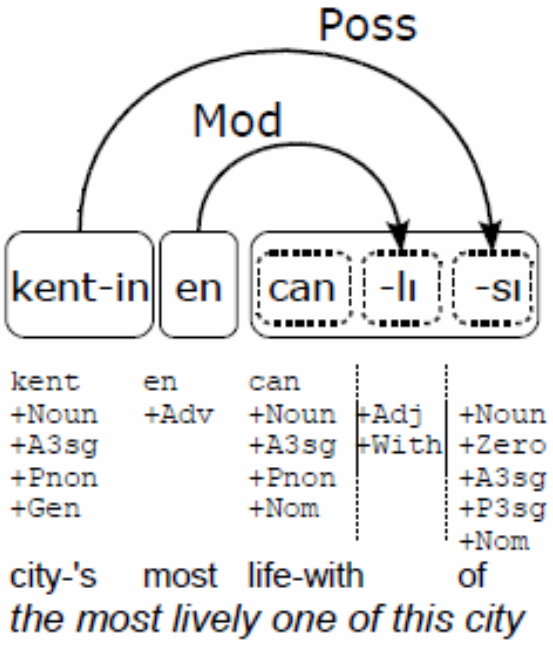
- Functional Structure
  - Sets of attribute value pairs
  - Attributes are features like tense and gender, or functions like subject and object
  - Values can be simple or be subsidiary f-structures
  - Functions of phrases → Language “independent”

C-structure and F-structure

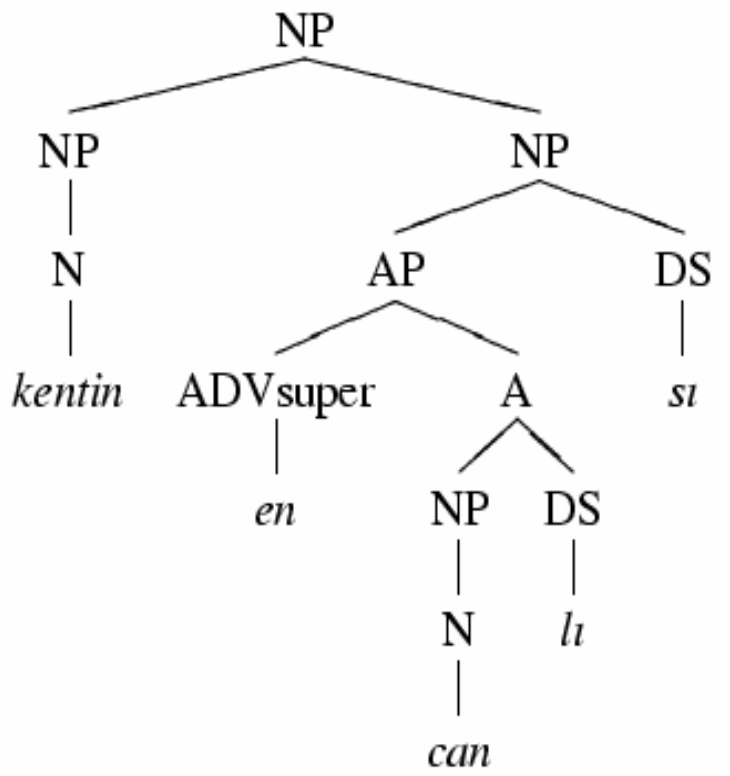


Inflectional Groups and Syntactic Relations

- Adverbial *en* modifies the derived adjective *canlı*
- The modified adjective is derived into a noun
- *kentin* (modifying *yeri* in the first example) modifies derived noun *canlısı*



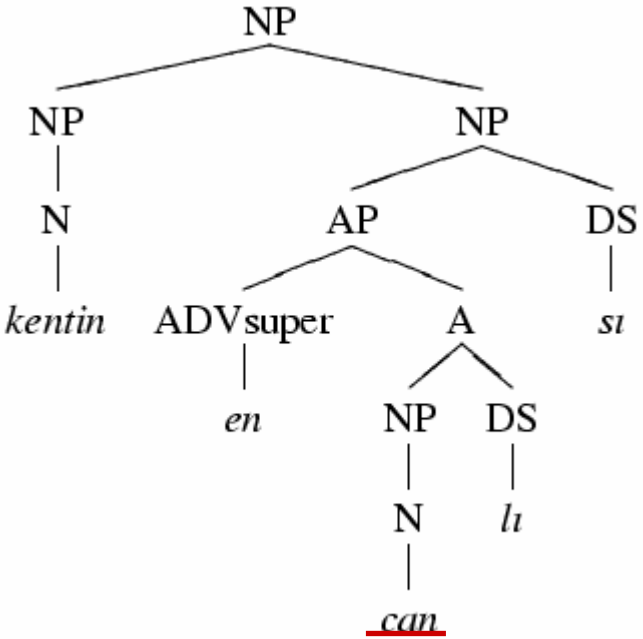
Inflectional Groups in LFG



- Each IG corresponds to a separate node in c-structure representation
- If an IG contains the root morpheme of the word, then the node corresponding to that IG is named as one of the syntactic category symbols
- The rest of the IGs are given the node name DS (to indicate derivational suffix)

The most lively one of the city

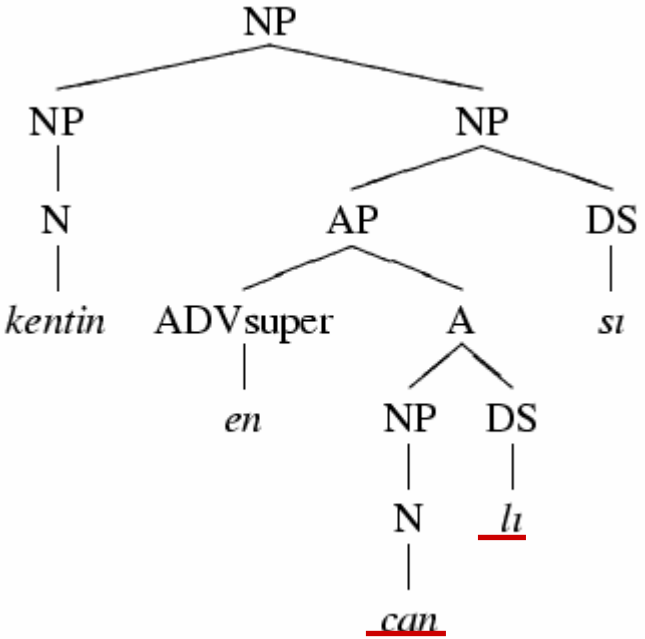
Inflectional Groups in LFG



$$\begin{bmatrix}
\text{PRED} & \text{'null-pro'} \\
\text{ADJUNCT} & \left\{ \begin{bmatrix}
\text{PRED} & \text{'li}\langle \text{CAN} \rangle\text{'} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'can'} \\ \text{CASE nom, NUM sg, PERS 3} \end{bmatrix} \\
\text{ADJUNCT} & \{ \begin{bmatrix} \text{PRED} & \text{'en'} \end{bmatrix} \} \\
\text{ATYPE attributive, DEGREE superlative}
\end{bmatrix} \right\} \\
\text{SPEC} & \begin{bmatrix} \text{POSS} & \begin{bmatrix} \text{PRED} & \text{'kent'} \\ \text{CASE gen, NUM sg, PERS 3} \end{bmatrix} \end{bmatrix} \\
\text{CASE NOM, NUM SG, PERS 3}
\end{bmatrix}$$

- Each node in c-structure corresponds to a separate f-structure
- the f-structure of the modifier is the value of an attribute in the f-structure of the head

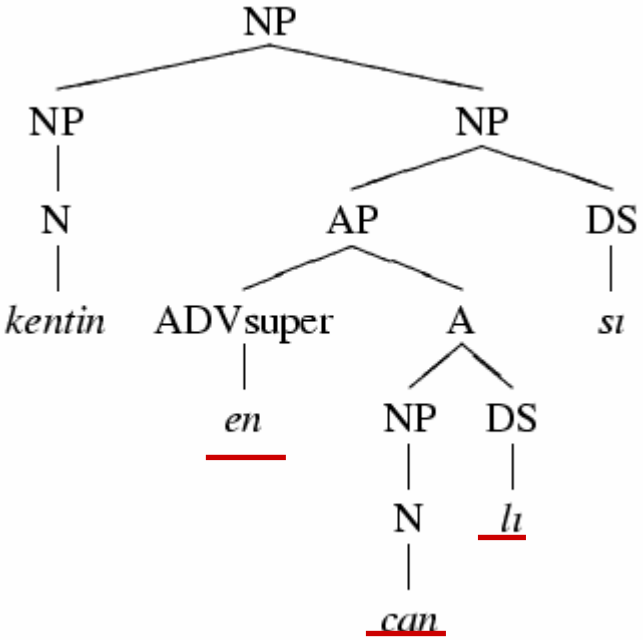
Inflectional Groups in LFG

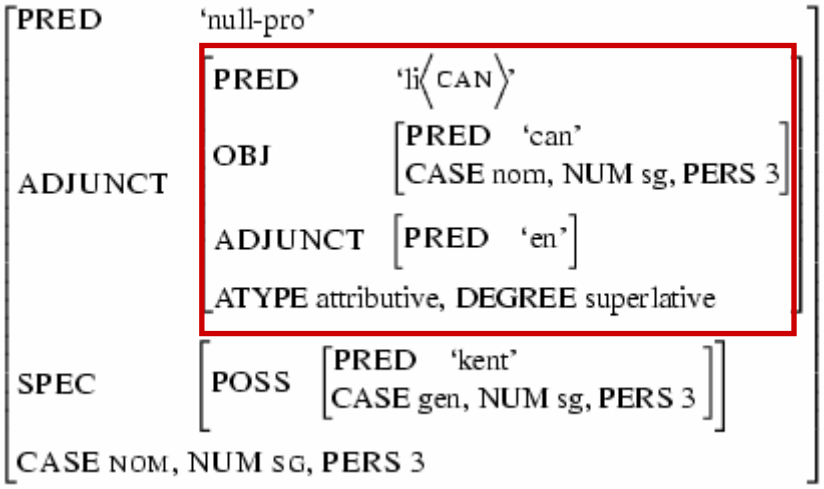


$$\begin{bmatrix}
\text{PRED} & \text{'null-pro'} \\
\text{ADJUNCT} & \left\{ \begin{bmatrix} \text{PRED} & \text{'lı}\langle \text{CAN} \rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'can'} \\ \text{CASE nom, NUM sg, PERS 3} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{ADJUNCT} & \begin{bmatrix} \text{PRED} & \text{'en'} \end{bmatrix} \\ \text{ATYPE attributive, DEGREE superlative} \end{bmatrix} \right\} \\
\text{SPEC} & \begin{bmatrix} \text{POSS} & \begin{bmatrix} \text{PRED} & \text{'kent'} \\ \text{CASE gen, NUM sg, PERS 3} \end{bmatrix} \end{bmatrix} \\
\text{CASE NOM, NUM SG, PERS 3}
\end{bmatrix}$$

- First, *can* (life) is derived into *canlı* (lively)
- NP → N
- A → NP DS

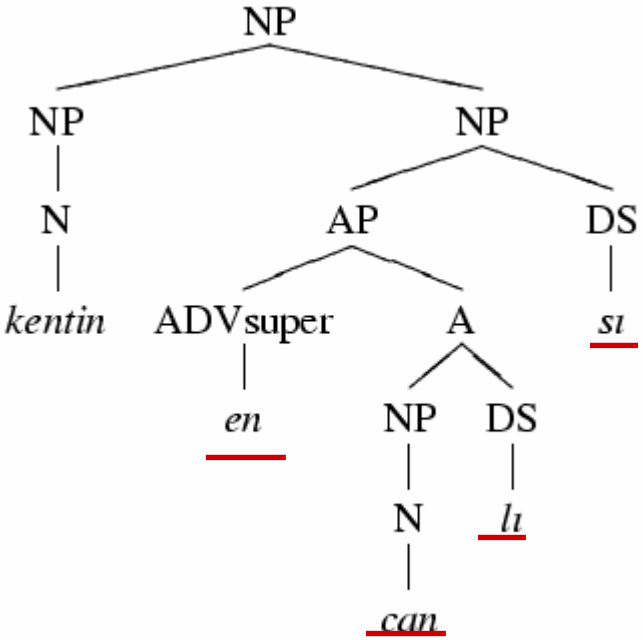
Inflectional Groups in LFG





- Then, superlative adverb *en* (most) modifies the adjective *canlı* (lively)
- AP → ADVsuper A

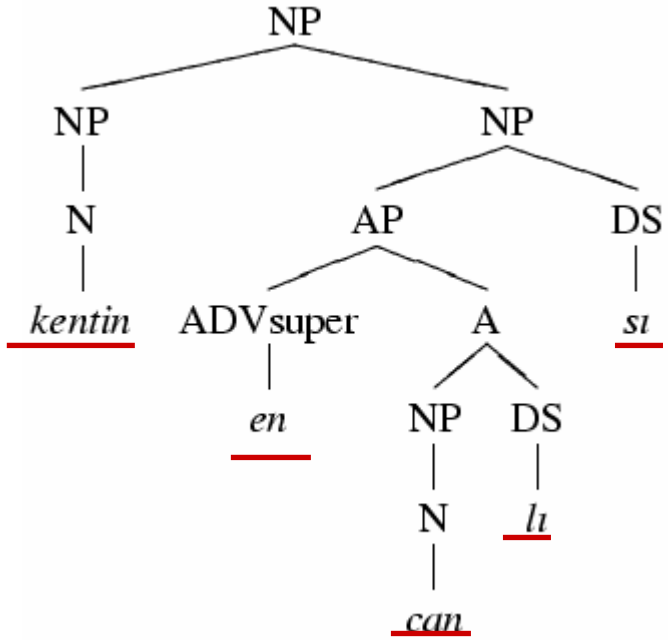
Inflectional Groups in LFG



$$\begin{bmatrix} \text{PRED} & \text{'null-pro'} \\ \text{ADJUNCT} & \begin{bmatrix} \text{PRED} & \text{'li}\langle \text{CAN} \rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'can'} \\ \text{CASE nom, NUM sg, PERS 3} \end{bmatrix} \\ \text{ADJUNCT} & \begin{bmatrix} \text{PRED} & \text{'en'} \end{bmatrix} \\ \text{ATYPE attributive, DEGREE superlative} \end{bmatrix} \\ \text{SPEC} & \begin{bmatrix} \text{POSS} & \begin{bmatrix} \text{PRED} & \text{'kent'} \\ \text{CASE gen, NUM sg, PERS 3} \end{bmatrix} \end{bmatrix} \\ \text{CASE NOM, NUM SG, PERS 3} \end{bmatrix}$$

- The whole AP *en canlı* (the most lively) is converted into an NP (the most lively one)
- No explicit derivational suffix
- NP → AP DS

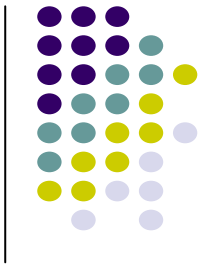
Inflectional Groups in LFG



$$\begin{bmatrix}
\text{PRED} & \text{'null-pro'} \\
\text{ADJUNCT} & \begin{bmatrix}
\text{PRED} & \text{'li}\langle \text{CAN} \rangle\text{'} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'can'} \\ \text{CASE nom, NUM sg, PERS 3} \end{bmatrix} \\
\text{ADJUNCT} & \begin{bmatrix} \text{PRED} & \text{'en'} \end{bmatrix} \\
\text{ATYPE attributive, DEGREE superlative}
\end{bmatrix} \\
\text{SPEC} & \begin{bmatrix} \text{POSS} & \begin{bmatrix} \text{PRED} & \text{'kent'} \\ \text{CASE gen, NUM sg, PERS 3} \end{bmatrix} \end{bmatrix} \\
\text{CASE NOM, NUM SG, PERS 3}
\end{bmatrix}$$

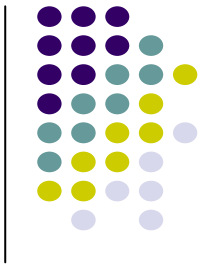
- NP *kentin* (of the city) specifies the NP *en canlısı* (the most lively one) as any usual NP
- NP → NP NP

Outline



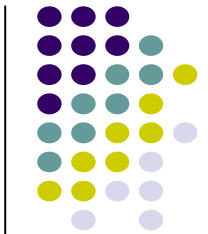
- Turkish in General
- Inflectional Groups
- Framework
- Work Accomplished
- Ongoing/Future Work
- Conclusion

Work Accomplished



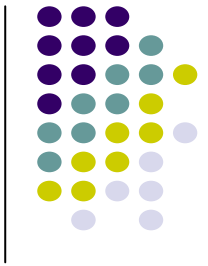
- Coverage
  - Noun phrases (definite, indefinite, pronoun,...)
  - Adjective phrases, adverbial phrases
  - Postpositions
  - Copular sentences
  - Basic sentences – free word order
  - Sentential derivations
  - Passives
  - Date-time expressions (Gümüş 2007)
- Linguistic Issues
  - Causatives
  - Non-canonical Objects

Sentential Derivations



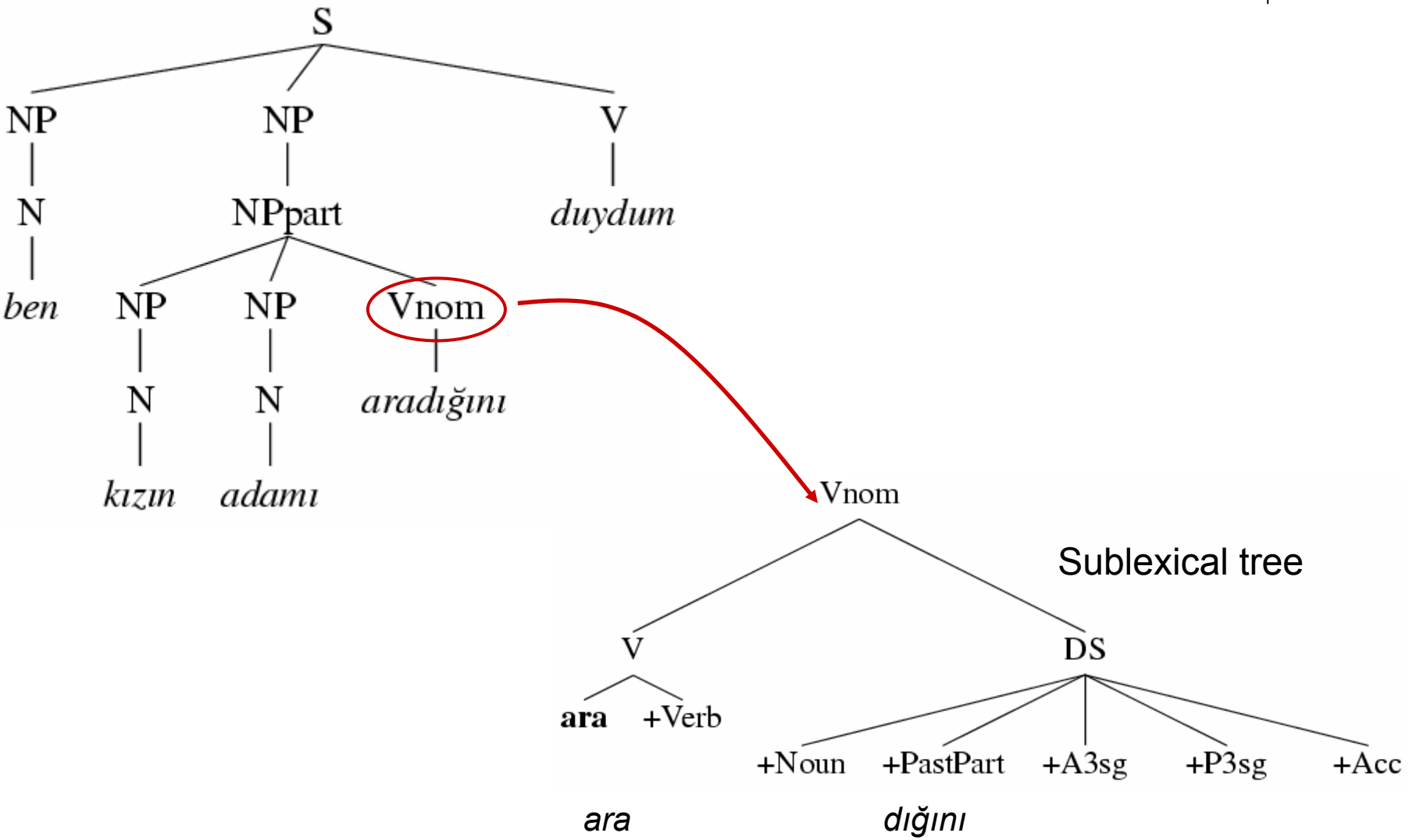
- Sentences can be used as constituents of other phrases by productive verbal derivations
- Sentences are derived into
  - Sentential complements
  - Participles
  - Adverbials
- Long distance dependencies in participles
  - Functional Uncertainty (Kaplan and Zaenen 1989)
  - regular expressions to define infinite path possibilities on one side of the constraints

Sentential Derivations

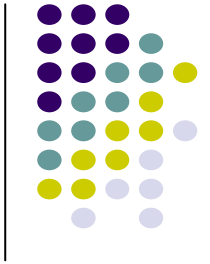


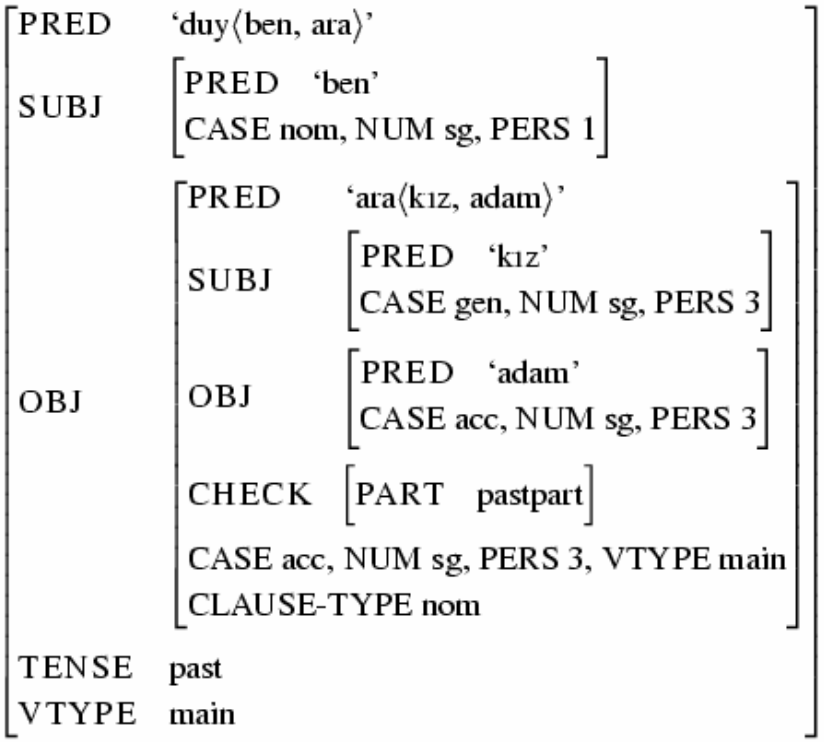
- kız adamı aradı. (the girl called the man)
  - ben kızın adamı aradığını duydum.
    I heard that the girl called the man.
  - $[\ ]_i$ adamı arayan $kız_i$
    the girl who calls the man
  - kız adamı ararken polis geldi.
    the police came while the girl called the man.

Sentential Complement C-structure



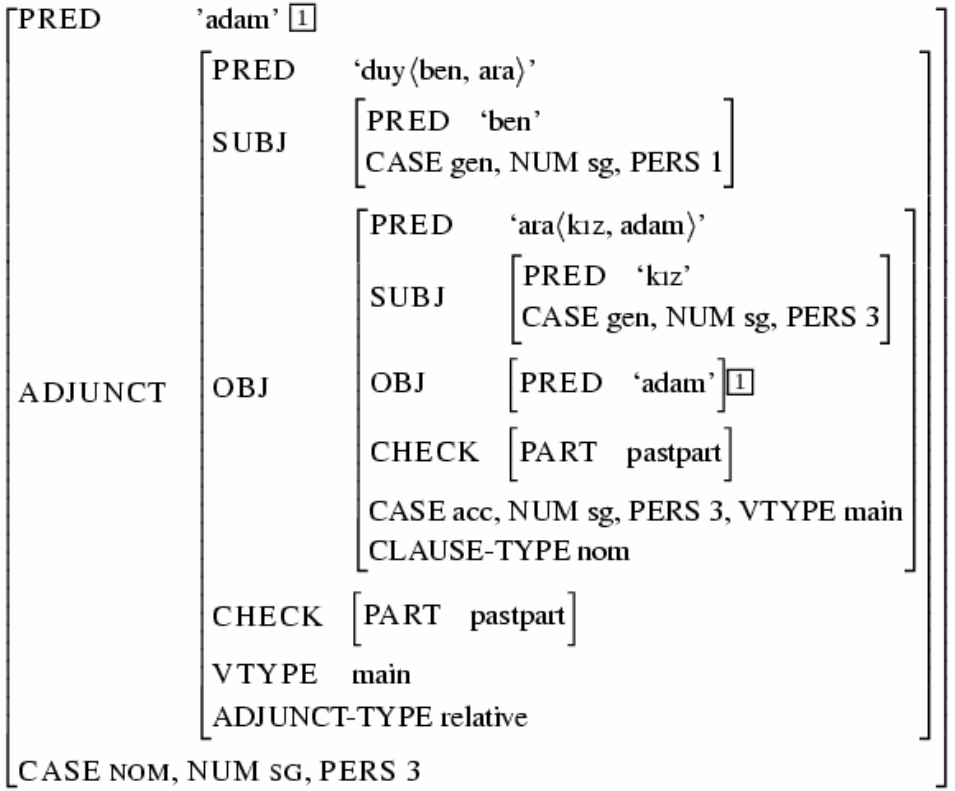
Sentential Derivations F-structure





ben kızın adamı aradığını duydum

(I heard the girl called the man)

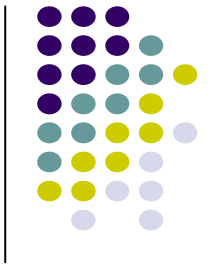


benim kızın [ ]$_i$ aradığını duyduğum adam$_i$

(the man I heard the girl called)

(↓ OBJ+) = ↑

Causatives



- Morphological process in Turkish

  aradı (s/he called)

  `ara+Verb+Pos+Past+A3sg`

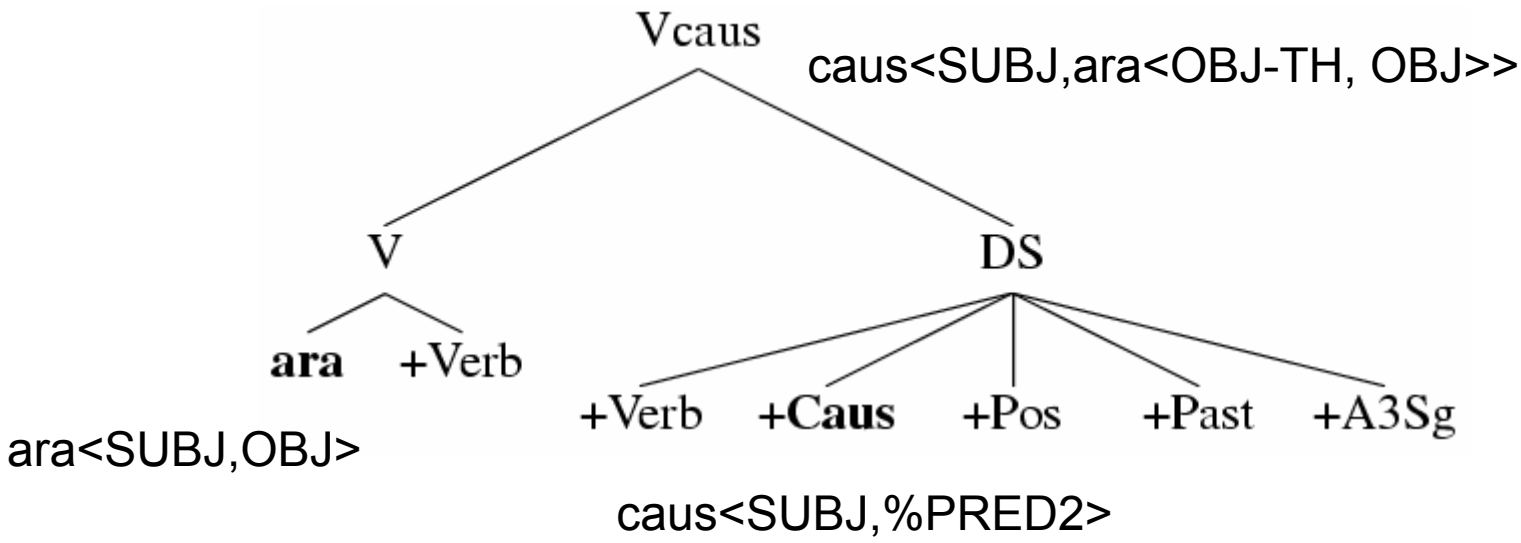
  arattı (s/he made her/him call)

  `ara+Verb^DB+Verb+Caus+Pos+Past+A3sg`

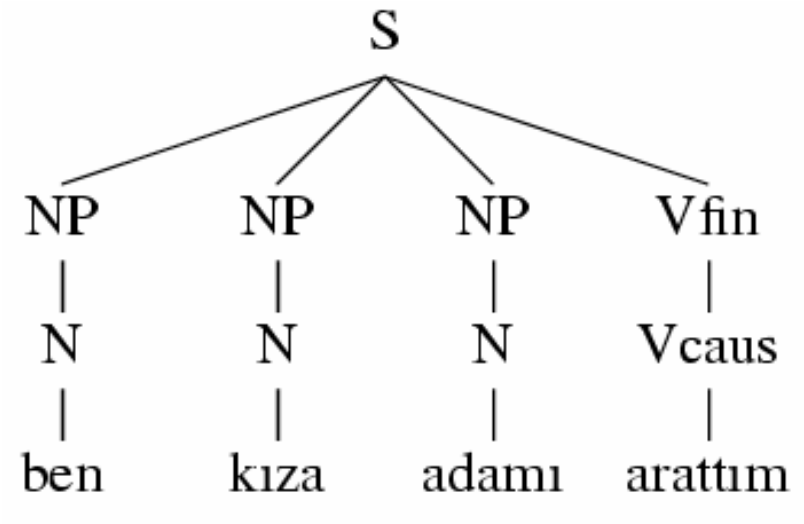
- How to represent?
  - with a single predicate (monoclausal) or with an embedded clause (biclausal)?
  - tests to identify the representation
  - details in (Çetinoğlu, Butt and Oflazer 2008)

Causative Implementation

- Two morphemes with predicative information: the verb stem and the causative morpheme
- These two predicates are merged to form a new complex predicate
- Following the approach in (Butt and King 2006)



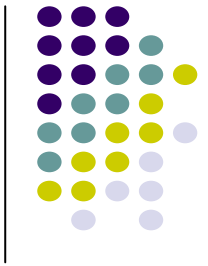
Causative C-structure



- Flat sentence structure to allow free order for all the constituents
- Case markers determine the functions of the phrases

(I made the girl call the man)

Causative F-structure



- The former nominative SUBJ becomes dative OBJ-TH
- Former OBJ in accusative case preserves its case and function

$$\begin{bmatrix} \text{PRED} & \text{'ara⟨kız,adam⟩'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'kız'} \\ \text{CASE} & \text{nom} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'adam'} \\ \text{CASE} & \text{acc} \end{bmatrix} \\ \text{TENSE} & \text{past} \\ \text{NUM SG, PERS 3, VTYPE MAIN} & \end{bmatrix}$$

kız adamı aradı
(the girl call the man)

$$\begin{bmatrix} \text{PRED} & \text{'caus⟨ben, ara⟨kız,adam⟩⟩'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'ben'} \\ \text{CASE} & \text{nom} \end{bmatrix} \\ \text{OBJ-TH} & \begin{bmatrix} \text{PRED} & \text{'kız'} \\ \text{CASE} & \text{dat} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'adam'} \\ \text{CASE} & \text{acc} \end{bmatrix} \\ \text{TENSE} & \text{past} \\ \text{NUM SG, PERS 1, VTYPE MAIN} & \end{bmatrix}$$

ben kıza adamı arattım
(I made the girl call the man)

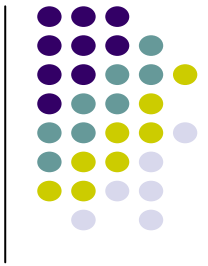
Non-canonical Objects

- Dative or ablative objects
- Can be divided into four main subgroups
- Have different causativization and passivization behavior
  - Studied and solution proposed in (Çetinoğlu and Butt 2008)

Hasan ata bindi (Hasan rode the horse)

Babası Hasan'ı ata bindirdi (His father made Hasan ride the horse)

Non-canonical Objects F-structures



- *bin* (ride) subcategorizes for SUBJ and OBJTH
- When causativized, former nom. SUBJ becomes acc. OBJ. OBJTH preserves its case and function

$$\begin{bmatrix} \text{PRED} & \text{'bin}\langle \text{Hasan, at}\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Hasan'} \\ \text{CASE} & \text{nom} \end{bmatrix} \\ \text{OBJTH} & \begin{bmatrix} \text{PRED} & \text{'at'} \\ \text{CASE} & \text{dat} \end{bmatrix} \\ \text{TENSE} & \text{PAST} \end{bmatrix}$$

Hasan ata bindi
(Hasan rode the horse)

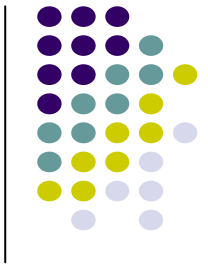
$$\begin{bmatrix} \text{PRED} & \text{'caus}\langle \text{baba, bin}\langle \text{Hasan, at}\rangle\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'baba'} \\ \text{CASE} & \text{nom} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'Hasan'} \\ \text{CASE} & \text{acc} \end{bmatrix} \\ \text{OBJTH} & \begin{bmatrix} \text{PRED} & \text{'at'} \\ \text{CASE} & \text{dat} \end{bmatrix} \\ \text{TENSE} & \text{PAST} \end{bmatrix}$$

Babası Hasan'ı ata bindirdi
(His fatherHasan ride the horse)

Related Issues

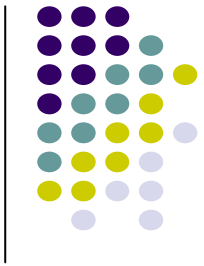
- Double causatives
  - Intransitives: similar to single causativization of transitives
  - Transitives: one of the arguments of the predicate is never explicit in the sentence
- Passivization
  - Basic, impersonal, double
  - Passivization of causatives
- Noun-verb complex predicates
  - yardım etmek (help), tamir etmek (repair), acı çekmek (suffer)

Outline



- Turkish in General
- Inflectional Groups
- Framework
- Work Accomplished
- Ongoing/Future Work
- Conclusion

Coordination

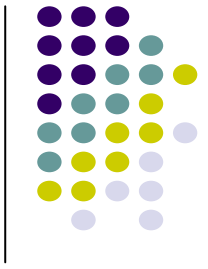


- Important in terms of coverage and performance
- Suspended Affixation (Kabak 2007)
  - All other coordinated constituents have certain default features which are then “overridden” by the features of the last element in the coordination

- kedilerden ve köpeklerden
- [kedi ve köpek]lerden (from cats and dogs)

- çalışırdık ve başarırdık
- [çalışır ve başarır]dık (we used to work and succeed)

Optimal Solutions



Kimse bana bu kötü büyüyü bozacak sihirli sözcüğü fısıldayamadı

(Nobody was able to whisper me the magical word that will break this bad spell)

- kimse : 1. nobody 2. person
- bana: 1. to me 2. to the “ban” (Ottoman title for Crotian princes)

- bu kötü büyüyü bozacak sihirli sözcüğü

bu kötü büyüyü bozacak sihirli sözcüğü

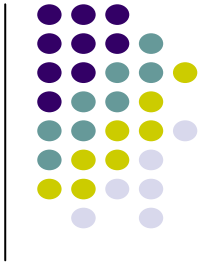
Optimal Solutions

Kimse bana bu kötü büyüyü bozacak sihirli sözcüğü fısıldayamadı

(Nobody was able to whisper me the magical word that will break this bad spell)

- OT-Marks (Frank et.al 2001)
  - Optimality Theory (Prince and Smolensky 2004) is applied for disambiguation by using OT-marks
  - Rules that cause a phrase to have different parses are marked with OT-marks
  - Then those marks are ranked in a user defined order

Testing



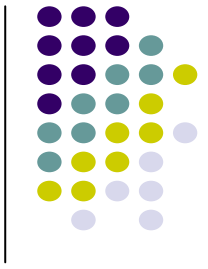
- Manual test files (~400)
- ParGram sentences (110)
- Tübitak progress report sentence test (43)
- Tübitak progress report noun phrase test (297)
  - Two random files from METU Corpus (Say et.al. 2002)
  - NPs manually extracted and grouped

TYPE	NUMBER	PARSED
Basic	194	182
Participle	48	37
Sentential	36	30
Coordination	19	5
Total	297	254 (85,5%)

Integrating LFG Grammar with LingBrowser

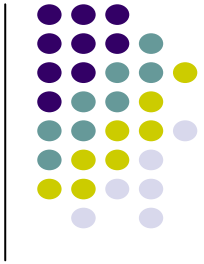
- LingBrowser (Armağan 2008)
  - NLP based browser for linguistic information
  - Word frequencies, morphological analysis, ...
  - Implemented as a Firefox add-on in Java
  - LFG parser available in the right click menu
    - pops up XLE-Web interface (Paul Meuer, University of Bergen)

Conclusion



- Building a large scale grammar is time consuming and linguistically challenging
- Coverage is one of the primary concerns
  - the tasks of performance criteria are accomplished
  - Naturally, the linguistic concerns are not ignored
  - but implementation of some infrequent usages or exceptional cases is eliminated

Publications



- Özlem Çetinoğlu and Kemal Oflazer, *Integrating Derivational Morphology into Syntax*, invited chapter in N. Nicolov et al.(eds.) Recent Advances in Natural Language Processing V: Amsterdam, John Benjamins, to appear in 2009.

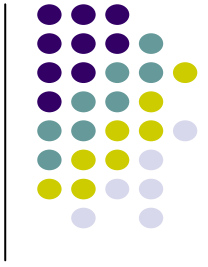
- Özlem Çetinoğlu, Miriam Butt, Kemal Oflazer, *Mono/Bi-clausality of Turkish Causatives*, International Conference on Turkish Linguistics, Antalya, August 2008.

- Özlem Çetinoğlu and Miriam Butt, *Turkish Non-canonical Objects*, in Proceedings of LFG'08 Conference, Sydney, Australia, July 2008.

- Özlem Çetinoğlu and Kemal Oflazer, *Morphology-Syntax Interface for Turkish LFG*, in Proceedings of COLING/ACL 2006, Sydney, Australia, July 2006

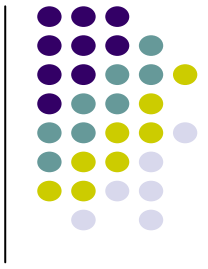
- Özlem Çetinoglu and Kemal Oflazer, *Altsözcüksel Birimlerle Türkçe için Sözcüksel İşlevsel Gramer Geliştirilmesi* [in Turkish], in Proceedings of the Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN 2006), Gökova, Muğla, June 2006

Thanks



?

Previous Work



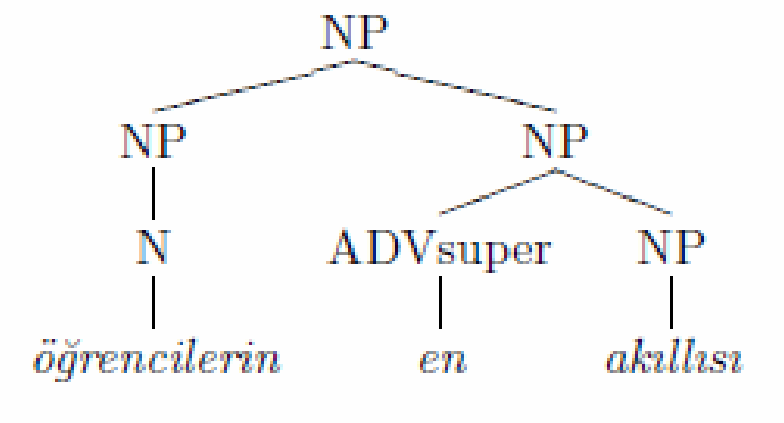
- HPSG (Şehitoğlu 1996)
- Categorial Grammar (Hoffman 1995)
- Principles and Parameters (Birtürk 1998)
- Combinatory Categorial Grammar (Bozşahin 2002)

- LFG (Güngördü and Oflazer 1995)

- Dependency Parser (Eryiğit and Oflazer, 2003)
- CCG (Çakıcı 2005)

Lexical Integrity

- Bresnan and Mugane 2006
  - Every lexical head is a morphologically complete word formed out of different elements and by different principles from syntactic phrases.



- 5 tests in (Bresnan and Mchombo 1995), 3 of them applicable for Turkish

Lexical Integrity

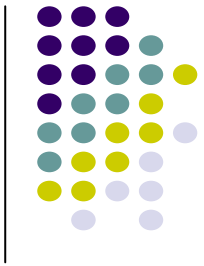
- Conjoinability

  ...while syntactic categories can be conjoined by syntactic conjunctions, stems and affixes normally cannot...

a. *gençtir* *ve* *güzeldir*
young-COP and beautiful-COP
s/he is young and beautiful

b. *[genç* *ve* *güzel]dir*
[young and beautiful]-COP

Lexical Integrity



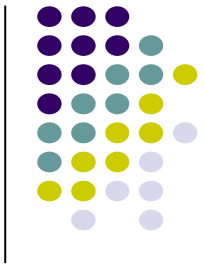
- Inbound Anaphoric Islands

  ...while phrases can contain anaphoric and deictic uses of syntactically independent pronouns, derived words and compounds cannot...

a. [kedi]siz (without a cat)
b. [on]suz (without it)

a. [Ali'de]ki (the one at Ali)
b. [onda]ki (the one at him)

Lexical Integrity



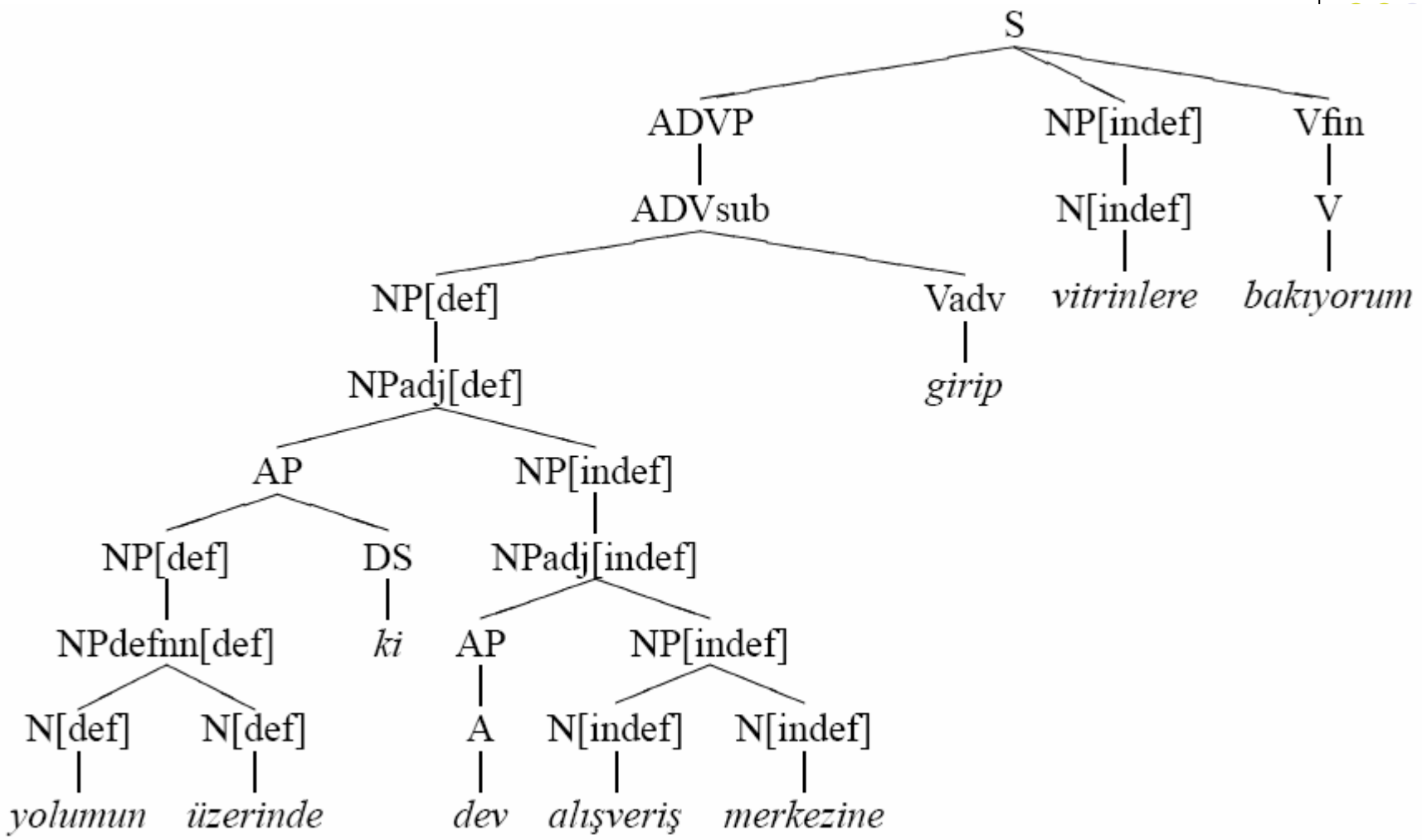
- Phrasal Recursivity

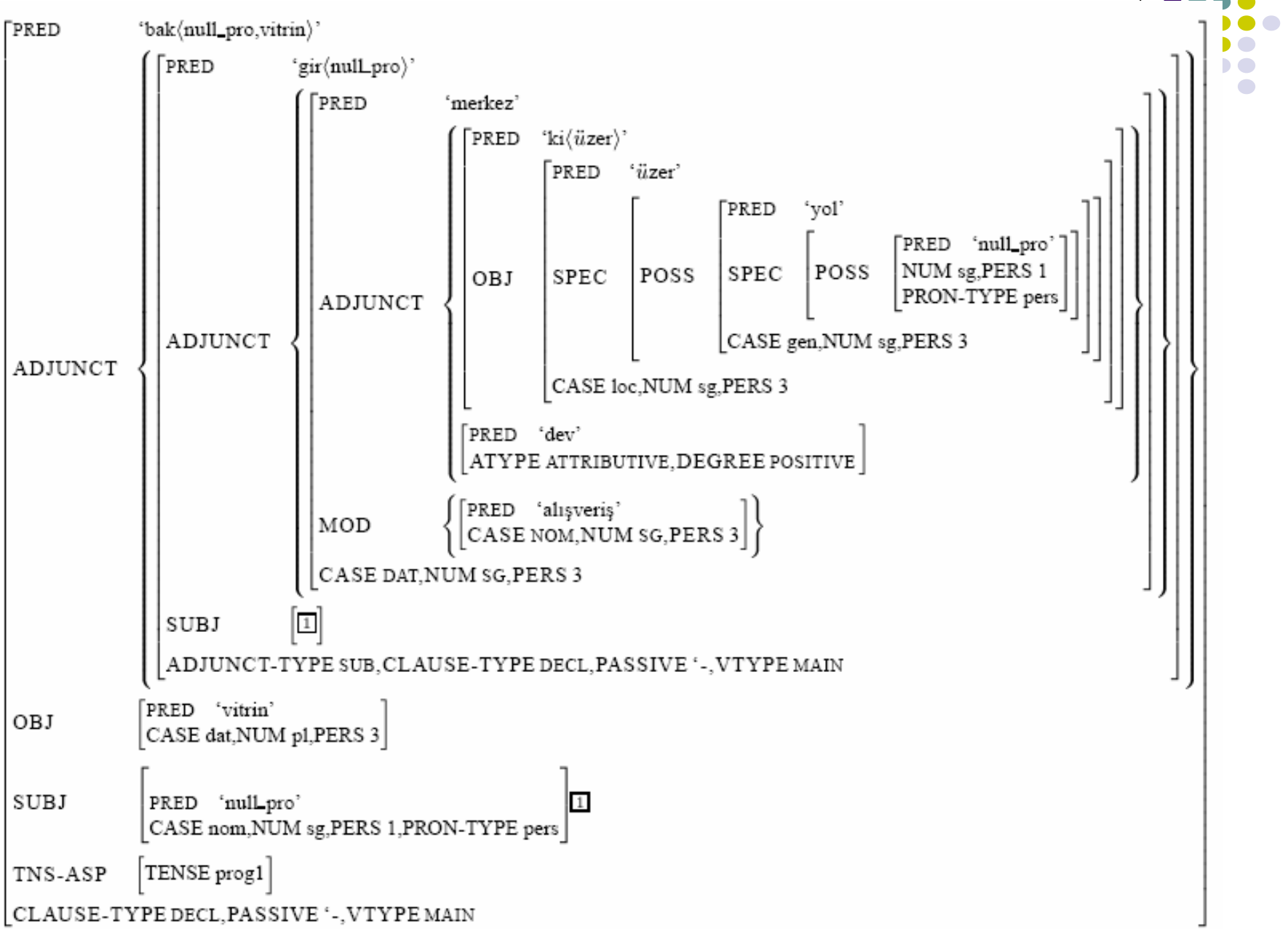
  ...word-internal constituents generally differ from word-external phrases in disallowing the arbitrarily deep embedding of syntactic phrasal modifiers...

a. *evde-ki*
house-LOC-REL
in the house

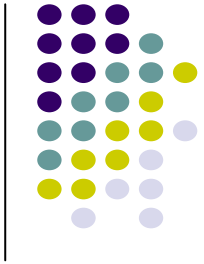
b. *[bu evde]ki*
[this house-LOC]-REL
in this house

c. *[senin evinden daha güzel evde]ki*
[your ev-POSS-ABL more beautiful house-LOC]-REL
in the house which is more beautiful than your house


S
ADVP
NP[indef]
Vfin
ADVsub
N[indef]
V
NP[def]
Vadv
vitrinlere
bakıyorum
NPadj[def]
girip
AP
NP[indef]
NP[def]
DS
NPadj[indef]
NPdefnn[def]
ki
AP
NP[indef]
N[def]
N[def]
A
N[indef]
N[indef]
yolumun
üzerinde
dev
alışveriş
merkezine


PRED 'bak⟨null_pro,vitrin⟩'
ADJUNCT
PRED 'gir⟨null_pro⟩'
ADJUNCT
PRED 'merkez'
ADJUNCT
PRED 'ki⟨üzer⟩'
OBJ
PRED 'üzer'
SPEC
POSS
PRED 'yol'
SPEC
POSS
PRED 'null_pro'
NUM sg,PERS 1
PRON-TYPE pers
CASE gen,NUM sg,PERS 3
CASE loc,NUM sg,PERS 3
PRED 'dev'
ATYPE ATTRIBUTIVE,DEGREE POSITIVE
MOD
PRED 'alışveriş'
CASE NOM,NUM SG,PERS 3
CASE DAT,NUM SG,PERS 3
SUBJ 1
ADJUNCT-TYPE SUB,CLAUSE-TYPE DECL,PASSIVE '-,VTYPE MAIN
OBJ
PRED 'vitrin'
CASE dat,NUM pl,PERS 3
SUBJ
PRED 'null_pro'
CASE nom,NUM sg,PERS 1,PRON-TYPE pers
1
TNS-ASP
TENSE prog1
CLAUSE-TYPE DECL,PASSIVE '-,VTYPE MAIN

Facts and Figures



English	#Rules	#States	#Disjuncts
Turkish	418	14526	69332
	103	1998	15755

Time in CPU seconds

TYPE	Total	Max
Basic	8.42	2.98
Participle	33.21	4.62
Sentential	12.11	4.28
Coordination	0.93	0.43