

FAST MORPHOLOGICAL ANALYSIS OF AGGLUTINATIVE LANGUAGES - APPLICATION ON TURKISH -

March 26-30, 2001

NATO IST

PARIS

M. OĞUZHAN KÜLEKÇİ

Senior Researcher

National Research Institute of Electronics & Cryptology

kulekci@uekae.tubitak.gov.tr

MEHMED ÖZKAN

Assoc. Prof.

Boğaziçi University

Biomedical Engineering Institute

İSTANBUL

mehmed@boun.edu.tr

OUTLINE

- Some Background
- Description
- Case Study :
 - Implementation of the technique on Turkish Language.
- Conclusions
- Future work

NLP

Phonetic Level

The compositions of speech sounds.

Morphological Level

The form of words in a sentences;
Verb Conjunctions; Plural of common nouns etc.

Derivational Morphology

word formation by affixes, appendixes
from other words, nominalizations of verbs

Syntactic Level

The way words combine to form sentences.

Semantic Level

Meanings of words, word groups and sentences.

Pragmatic Level

Relationship between an utterance (or sentence) and its socio-cultural context.

- In Turkish, approximately 50% of unsuccessful search operations are due to the morphological problems related with the words being searched.
- Search engines use natural language processing tools to overcome that problem (as in AltaVista Search Engine 3.0 for some languages).
- The pillar of an agglutinative language NLP is the morphological analyzer.
- The performance of higher levels (syntax, semantics) are limited by the power of the morphological analyzer.
- The morphological analyzer used in a search engine must be fast, robust and easy to update.

MA

- **There is as yet no standard approach to computational morphology**

(Covington, NLP for Prolog Programmers, Prentice Hall).

- In general, finite state or statistical methods are being used in the morphological analyzers.
 - * Finite State Transition Networks (FSTN)
 - * Finite State Transducers (FST)
 - * Recursive Transition Networks (RTN)
 - * Augmented Transition Networks (ATN)
 - * Two-Level Morphology-KIMMO
(Koskenniemi - 1993) : An alternate way of implementation to handle the underlying and surface form differentiation.
- A morpheme group search technique is proposed for agglutinative languages.

KOZ

KOZ does the analysis of words by recognizing morphemes in **GROUPS!**

Sample word:

Evlerinden (from their houses)

Morphological analysis of a word with the classical methodologies:
(suffixes are detected one by one)

Ev – ler – in - den

Morphological analysis of the same word with KOZ:
(suffixes are detected in groups)

Ev - lerinden

KOZ

Morpheme groups must be generated
and compiled into databases by
preprocessing.

Definitions:

- **Allomorph-Set** : Morphemes that are in the same grammar classification and have the same morphophonemic characteristic of the language form an **allomorph-set**.

Example:

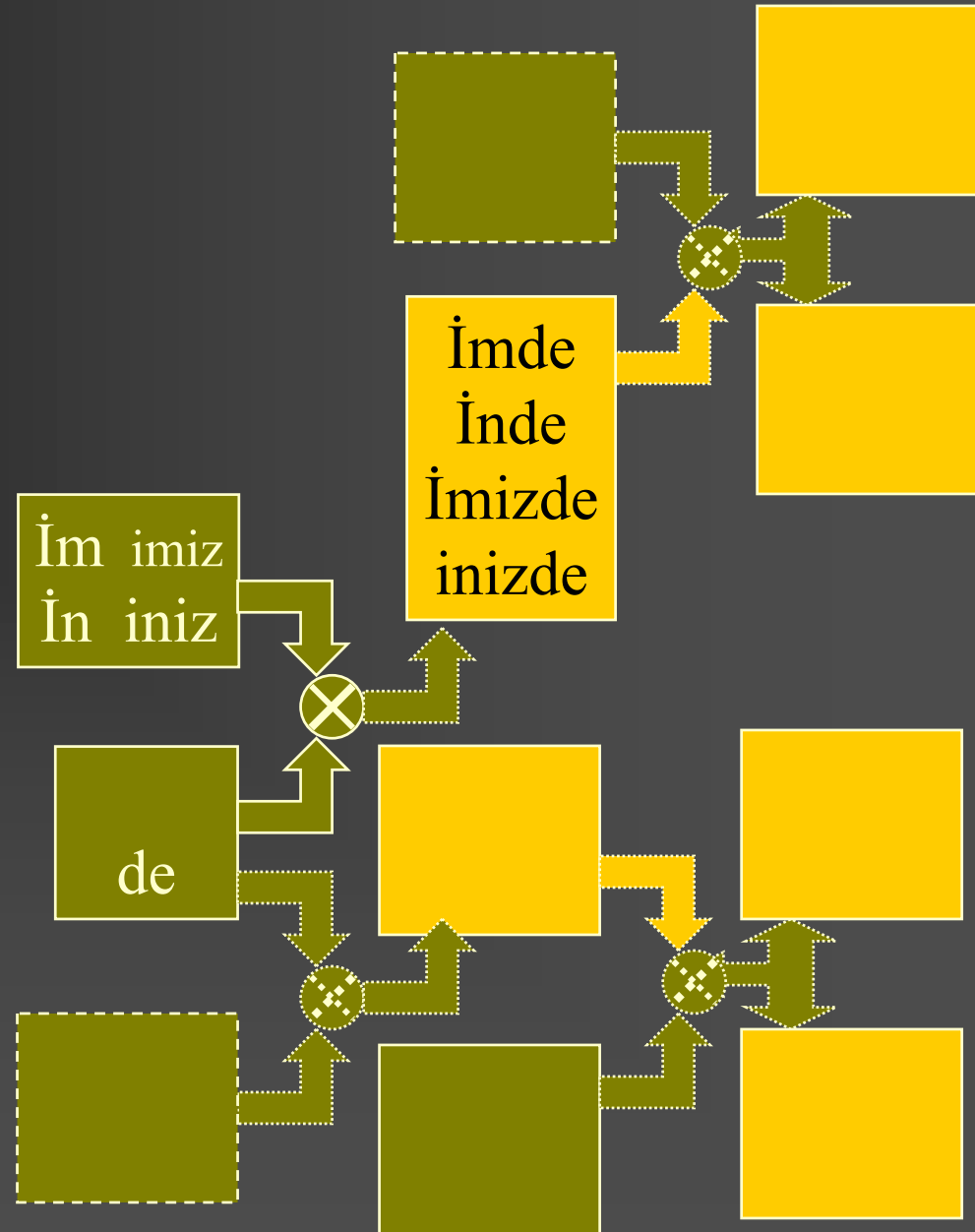
Two sets of allomorphs of Turkish possessive morpheme class differs by morphophonemic characteristics are:

{*im, in, imiz, iniz*} and {*ım, ın, ımız, ınız*}

.... similarly locative allomorph sets:

{*de*} and {*da*}

- **Morpheme-Group:**
The concatenation of allomorph-sets defined by morphotactic rules forms an morpheme-group.



Noun Inflection Structure in Turkish

Morphotactic rules specify the order of affixation.

PLURALITY	POSSESSIVE	CASE
0	0	0
0	0	Genitive
0	0	Accusative
0	0	Dative
0	0	Ablative
0	0	Locative
0	0	Instrumental
0	0	Equality
0	1	0
0	1	Genitive
0	1	Accusative
0	1	Dative
0	1	Ablative
0	1	Locative
0	1	Instrumental
0	1	Equality
1	0	0
1	0	Genitive
1	0	Accusative
1	0	Dative
1	0	Ablative
1	0	Locative
1	0	Instrumental
1	0	Equality
1	1	0
1	1	Genitive
1	1	Accusative
1	1	Dative
1	1	Ablative
1	1	Locative
1	1	Instrumental
1	1	Equality

Noun Inflection Allomorph-sets in Turkish

	plural	possessive	genitive	accusative	Dative	ablative	locative	instrumental	equality
1	ler	imiz,inizim,in	in	i	e	den	de	le	ca,casina
2	lar	ımız,ınızım,in	ın	ı	a	dan	da	le	nca,ncasina
3		umuz,unuz um,un	un	u	na	tan	ta	yla	ce,cesine
4		ümüz,ünüzüm,ün	ün	ü	ne	ten	te	yle	nce,ncesine
5		miz ,niz	nin	ni	ya	ndan	nda		ça,çasına
6		miz,niz	nin	ni	ye	nden	nde		çe,çesine
7		muz,nuz	nun	nu					layın
8		müz,nüz	nün	nü					leyin
9		m		yi					
10		n		yi					
11		mlar, nlar		yu					
12		mler, nler		yü					
13		si							
14		si							
15		u, su							
16		ü,sü							
17		i							
18		i							

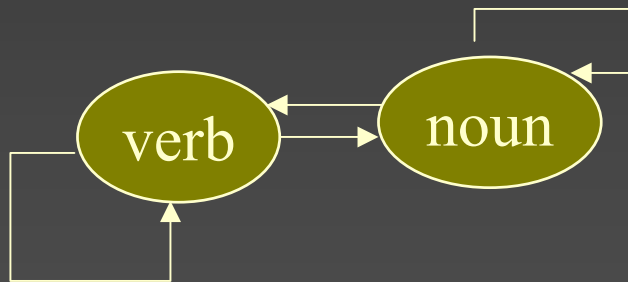
Morpheme-Group Generation Tables (MGGT) show all possible affixes

- Vowel/consonant harmony rules are investigated for morphophonemic classification of allomorphs

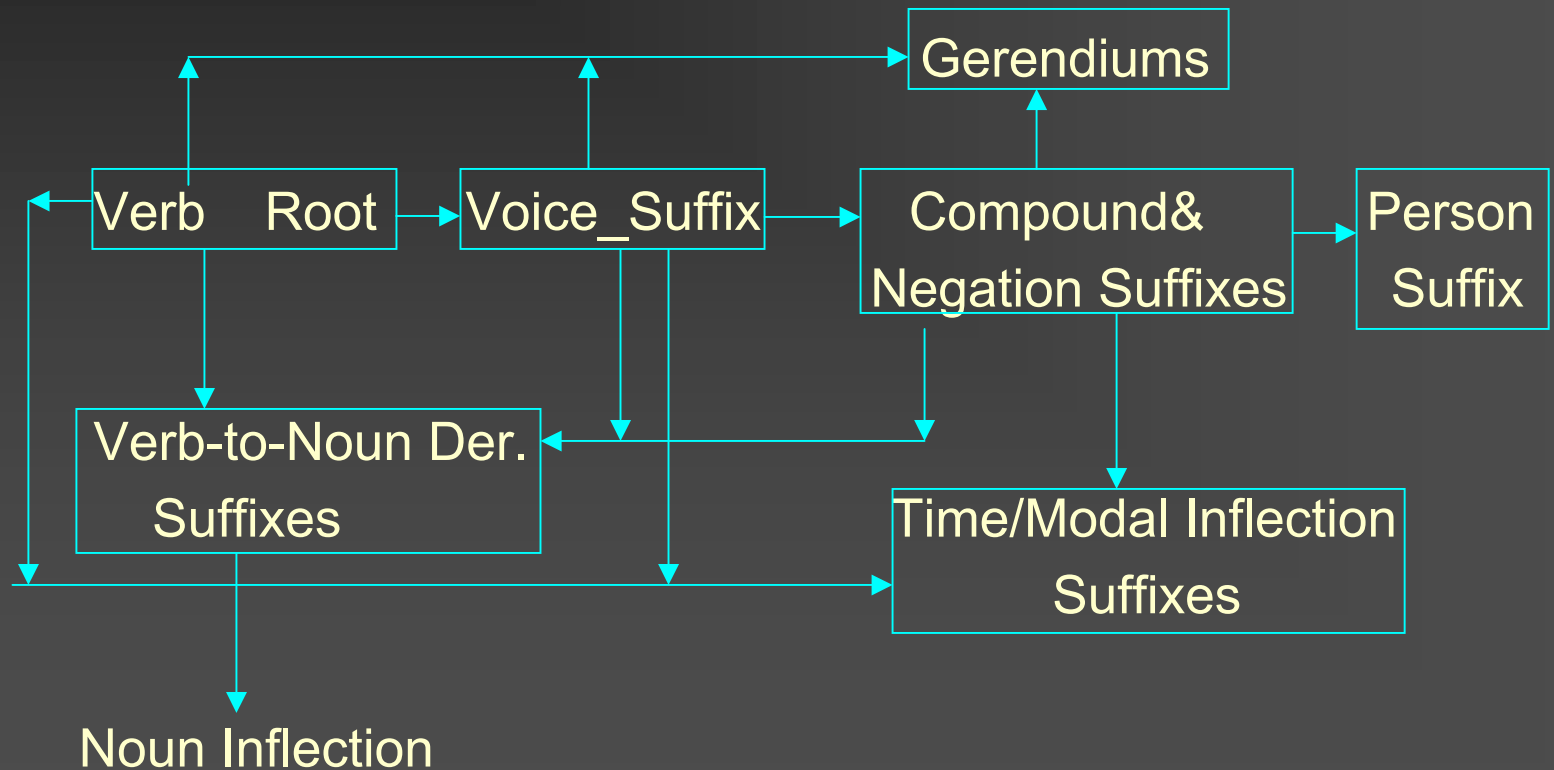
Plural	Possessive	Genitive	Accusative	Dative	Ablative	Locative	Instrumental	Equality
1	1	1	1	2	2	2	2	3
2	2	2	2	1	1	1	1	1
	3	3	3	1	1	1	1	1
	4	4	4	2	2	2	2	3
	5	2	2	1	1	1	1	1
	6	1	1	2	2	2	2	3
	7	3	3	1	1	1	1	1
	8	4	4	2	2	2	2	3
	9	1,2,3,4	1,2,3,4	1,2	1,2	1,2	1,2	1,3
	10	1,2,3,4	1,2,3,4	1,2	1,2	1,2	1,2	1,3
	11	2	2	1	1	1	1	1
	12	1	1	2	2	2	2	3
	13	6	6	3	5	5	3	2
	14	5	5	4	6	6	4	4
	15	7	7	3	5	5	3	2
	16	8	8	4	6	6	4	4
	17	6	6	3	5	5	3	2
	18	5	5	4	6	6	4	4

Inflection/derivation structure and word formation

Turkish has a recursive word derivation structure

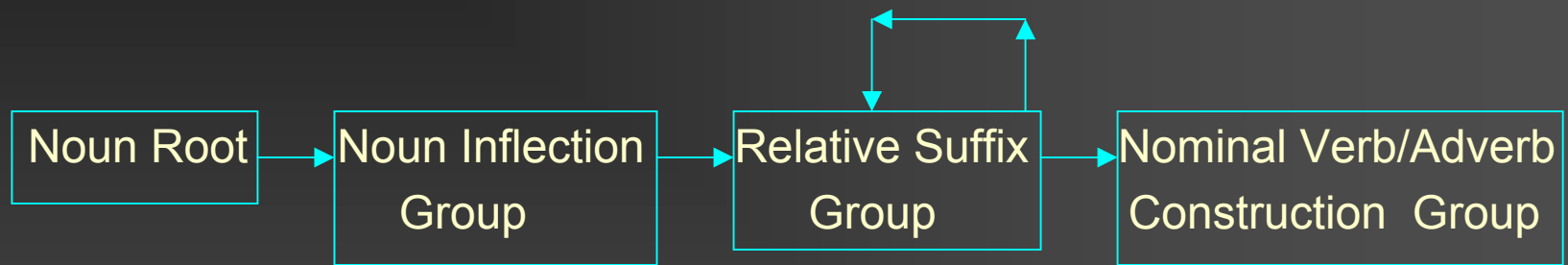


Turkish Verb Inflection



Turkish Noun Inflection

- Turkish Noun (substantive) inflection diagram



* Noun Inflection Group = Plural Suffixes + Possessive Suffixes + Case Suffixes

* Relative Suffix Group maybe recursive

Morpheme-Groups used in Turkish Morphological Analyzer

- 3184 Morpheme-Group(MG) for Verb Time/Modal Inflection
- 52 MG for Gerendiums
- 516 MG for Noun Inflection Group
- 296 MG for Nominal Verb/Adverb Constructions
- 74 MG for Relative ('ki' suffix) Group
- 59 MG for Person Suffix MG
- 272 MG for Noun derivation suffixes
- 6 MG for Verb derivation suffixes
- 604 MG for Voice-Suffixes
- 68 MG for Complementary&Negation Suffixes
- For the generation of MGs which are over 100 in quantity MGGTs are used.

With a total of 5131 morpheme groups, Turkish Morphology (a highly agglutinative and difficult grammar) is fully analyzed.

Summary

- **Definition of Morphophonemic Rules** : To determine phonemic constraints (vowel/consonant harmony rules) of morpheme concatenation.
- **Definition of Morphotactic Rules** : To determine morphotactic constraints (word formation, inflection/derivation rules) of morpheme concatenation.
- **Classification of morphemes and allomorph-set formations**
- **Preparation of necessary morpheme-group generation tables**
- **Generation of morpheme groups by MGGT**
- **Building s search engine.**

Conclusions

■ Advantages:

- enables the morpheme search for the morphological analysis of a word to be performed in groups instead of morpheme by morpheme search.
- Faster, good for real time application.
- No need for backtracking.
- Modifications and upgrades are practical and easy.

■ Disadvantage:

- Increase in memory space to store morpheme-groups. (500Kbyte.)

Future Work

- Better lexicon design
 - Exception handling
 - Speed optimization
 - Applications
 - Dictionary
 - Translation
 - ASR (language model)
 - Indexing for Internet
 - Keyword search
-

THANK YOU