Accommodating Multiword Expressions in an LFG Grammar

Mohammed A. Attia
The University of Manchester
School of Informatics
mohammed.attia@postgrad.manchester.ac.uk

The ParGram Meeting
Japan September 2005

Introduction

- Why handle MWEs?
  - MWEs are pervasive in all languages
    - they are in the same order of magnitude as the speaker's lexicon
    - they account for 41% of the entries in WordNet 1.7
  - Machine translation
    - Compositional Translation: Blind and literal
    - Non-compositional Translation: certain and accurate

Introduction

- Where to handle MWEs?
  - Not at higher phases of processing such as transfer
  - MWEs require deep analysis that starts as early as the normalization and tokenization, and goes through morphological analysis and into the syntactic rules

Introduction

- What is the advantage of handling MWEs?
  - Reduction of ambiguity
  - Avoiding needless analysis of idiosyncratic structures
  - Reduction of parsing time
  - Precise analysis

Definition

- Informally: A word with spaces
- Meaning units that cross word boundaries
- They cover
  - idioms (e.g. *down the drain*)
  - phrasal verbs (e.g. *rely on*)
  - verbs with particles (e.g. *give up*)
  - compound nouns (e.g. *book cover*)
  - collocations (e.g. *do a favour*)

Definition

- The term *multiword* itself has been challenged
  - a word as a string of letters between two delimiters
  - – There are languages that do not use spaces between words, such as Japanese.
  - – Compound nouns in German are written without spaces.
  - – Arabic has a group of clitics (pronouns, prepositions, definite article, etc.)

Definition

How to decide what expressions can be considered MWEs?

1. Lexogrammatical fixedness. The expression has come to a rigid or frozen state. The expression must be immune to the following operations:
   - Substitutability
     *many thanks -> * several thanks*
   - Deletion
     *black hole -> * the hole*
   - Category transformation
     *bitter cold -> * the bitterness of the cold*
   - Permutation
     *life guard -> * the guard of life*
     *kiss of life -> * life kiss*
     *day and night -> * night and day*

Definition

How to decide what expressions can be considered MWEs?

2. Semantic non-compositionality. The meaning of the expression is not driven from the meaning of the component parts.

   kick the bucket = die

Definition

How to decide what expressions can be considered MWEs?

3. Syntactic irregularity. The expression exhibits a structure that is inexplicable by regular grammatical rules.
   long time, no see
   by and large

Definition

How to decide what expressions can be considered MWEs?

4. Single-word paraphrasability. The expression can be paraphrased by a single word.
   give up = abandon

Definition

How to decide what expressions can be considered MWEs?

5. Single word translatability. Expressions can be considered as terms when
   - the corresponding translation is a unit
   - their translation differs from a word to word translation (Brun 1998).

Classification of Multiword Expressions

- Semantically
  - Compositional
  - Non-compositional
- Morphosyntactically
  - Flexible
  - Inflexible

Classification of Multiword Expressions

I. Compositional vs. Non-Compositional

- how the overall sense of a given idiom is related to its parts
- No binary distinction of plus and minus, free variety

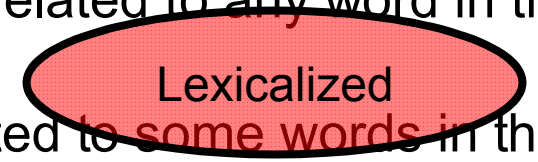
1. The meaning is not related to any word in the expression
   - *Kick the bucket*
2. The meaning is related to some words in the expression. One or more words are not used in the usual sense
   - *Kill time*
   - *Fall in love*
   - *Break the news*
3. The meaning is derived from all words in the expression
   - *Book cover*
   - *Health crisis*
   - *party meeting*

Classification of Multiword Expressions

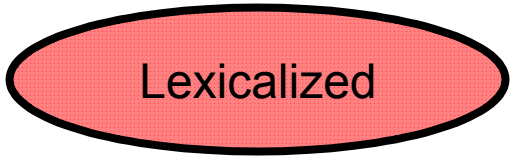
I. Compositional vs. Non-Compositional

- how the overall sense of a given idiom is related to its parts
- No binary distinction of plus and minus, free variety

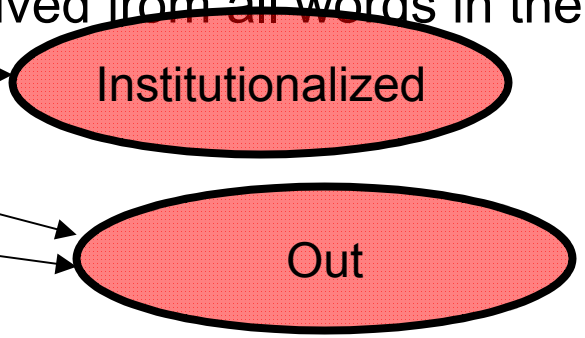
1. The meaning is not related to any word in the expression
   - *Kick the bucket*



2. The meaning is related to some words in the expression. One or more words are not used in the usual sense
   - *Kill time*
   - *Fall in love*
   - *Break the news*



3. The meaning is derived from all words in the expression
   - *Book cover*
   - *Health crisis*
   - *party meeting*



Classification of Multiword Expressions

I. Compositional vs. Non-Compositional

Collocations

words co-occur in a statically meaningful way

1. Frozen Modifiers
   - *Bitter cold*
   - *Scorching heat*
   - *Shining example*
   - حرب شعواء large-scale war
   - ظلام دامس gloomy darkness

Classification of Multiword Expressions

I. Compositional vs. Non-Compositional

Collocations

2. Support Verbs: selection is determined by the object noun
   - Class 1 (Light Verbs): no semantic value. Conjugate the meaning of the object noun
     - to have dinner
     - to get angry
     - to give attention
     - to play a part
     - to do harm
     - to make a request
     - to take care
   - Class 2: carry a semantic value. Used to express a scenario in the event
     - to fulfill a dream
     - to keep a promise
     - to pass an exam
     - to undergo an operation

Classification of Multiword Expressions

II. Flexible vs. Inflexible MWE

- Syntactic and morphological flexibility
  1. Fixed expressions: lexically, syntactically and morphologically rigid
     - *San Francisco*
     - *in a nutshel*

     Frozen Texts: expressions are frozen at the level of the sentence
     - Idiomatic (proverbial)
       - *A bird in hand is worth two in the bush*
       - *A friend in need is a friend indeed*
     - Pragmatic
       - *Good morning*
       - *We haven't got all day*

Classification of Multiword Expressions

II. Flexible vs. Inflexible MWE

2. **Semi-Fixed Expressions**: undergo morphological and lexical variations, but still the components are adjacent
   - Morphological variation
     - *traffic light/lights*
     - *kick/kicks/kicked the bucket*
   - Lexical variations
     - *to sweep something under the carpet/rug*
     - على وجه/ظهر الأرض/البسيطة
       on face/back the-earth/the-land
       on the face of the earth

Classification of Multiword Expressions

II. Flexible vs. Inflexible MWE

3. **Syntactically-Flexible Expressions**: expression that can either undergo reordering
   - passivization
     - *the cat was let out of the bag*
   - external elements intervene between the components
     - *give smoking up*
     - دراجة الولد البخارية
       bicycle the boy the fiery
       the boy's fiery bicycle
       the boy's motorbike

Extracting Multiword Expressions

- Electronic dictionaries:
  - Single words = available
  - MWEs = not as available
- Tools needed for automatic extraction
  - Tagger and/or Parser
    Pattern matching NN or AN or NPN
  - Corpus of translated texts

Extracting Multiword Expressions

- Manual Extraction
  - MWEs are added as you come across them
- Semi-automatic extraction
  - A list of terms that frequently occur as part of a MWE is built
    republic, kingdom, organization, council
  - These terms are tracked in a concordance tool
  - The output is sorted and filtered

Extracting Multiword Expressions

- Semi-automatic extraction of Arabic compound names and adverbs
  - Compound proper names usually have one of the following words as the initial component:
    - عبد Abd (slave of – compounded with one of the 99 attributes of Allah)
      - عبد الجواد Abd al-Jawwad (Lit. servant of the Generous)
      - عبد الرحمن Abd al-Rahman (Lit. servant of the Merciful)
    - بن Bin (son of)
      - بن لادن Bin Laden (Lit. son of Laden)
    - أبو Abu (father of)
      - أبو عمار Abu Ammar (Lit. father of Ammar)
  - Adverbs of manners
    - بطريقة (in a way) + adjective
      - بطريقة قانونية in a way legal (legally)
    - بشكل (in a form) + adjective
      - بشكل قاطع in a form absolute (absolutely)

Handling MWEs

- Fixed expressions => Morphology
- Lexically-flexible expressions => Morphology
- Morphologically-flexible expressions => Morphology
- Syntactically-flexible expressions => Syntax

Handling MWEs

- **Building the MWE Transducer**
  - Finite state regular expression
  - Two-sided transducer is for MWEs
    - Analysis on the lexical side (upper side)
    - Generation on the surface side (lower side)
  - Fixed and semi-fixed expressions
  - Consults the core morphological transducer to account for the morphological flexibility

Handling MWEs

- **Building the MWE Transducer**: Implementation
  - load ArabicTransducer.fst
  - define AllWords
  - $[?* "[" {minister} "]" ?*] .o. AllWords
    sp ("+def":{the}) {foreign}
  - ["+noun" "+masc" "+def"]:{keeping}
    sp {peace}
  - $[?* "[" {car} "]" ?*] .o. AllWords
    sp $[?* "[" {trapping} "]" ?*] .o. AllWords

Handling MWEs

- **Building the MWE Transducer**: Combinatorial Rules
  - To filter out ungrammatical combination of words due to overgeneration

~$["+dual" <> ["+sg" | "+pl"] /?*]

.o. ~$["+fem" <> "+masc" /?*]

Handling MWEs

- **Building the MWE Transducer**: Feature Unification
  - To filter out repeated and redundant features from the analysis

    +noun+fem+sg[car] +adj+fem+sg[trapping]

    car trapping

  - We remove all features from the non-head element

    "+sg" -> [] || sp ?* _

    .o. "+fem" -> [] || sp ?* _

    .o. "+adj" -> [] || sp ?* _

    .o. "+noun" -> [] || sp ?* _

Handling MWEs

- Interaction with the tokenizer
  - MWEs composed with the tokenizer

```
regex [singleTokens.i .o.
?* 0:"[[[" (MweTokens.l) 0:"]]]" ?* .o.
"@" -> " " || "[[[" [Alphabet* | "@"*] _ [Alphabet* | "@"*] "]]]".o.
"[[[" -> [] .o.
"]]]" -> []].i;
```

Handling MWEs

- Interaction with the tokenizer

Input:

ولوزير خارجيتها
wa-liwazir kharijiyatiha
and-to-foreign minister-its
(and to its foreign minister)

Single Token Output:

و@ل@خارجية@وزير@ها@
(approx. and@to@ foreign@minister@its@)

MWEs and Final Output:

و@ل@خارجية وزير@ها
(approx. and@to@foreign minister@its@)

Handling MWEs

- Interaction with the white-space normalizer
  - Spaces are crucial in determining MWEs

No Space Before	No Space After
)	(
}	{
]	]
”	“
’	‘

Handling MWEs

- Interaction with the white-space normalizer
  - Input
    - وقال الولد ،لم أذهب(أو أمر بجوار)المدرسة .
  - Output
    - وقال الولد، لم أذهب (أو أمر بجوار) المدرسة.

Handling MWEs

- **Integration with the Morphological Transducer**
  - Union

Handling MWEs

- **Interaction with the grammar**
  - OT mark

أكل جنود حفظ الأمن التفاح

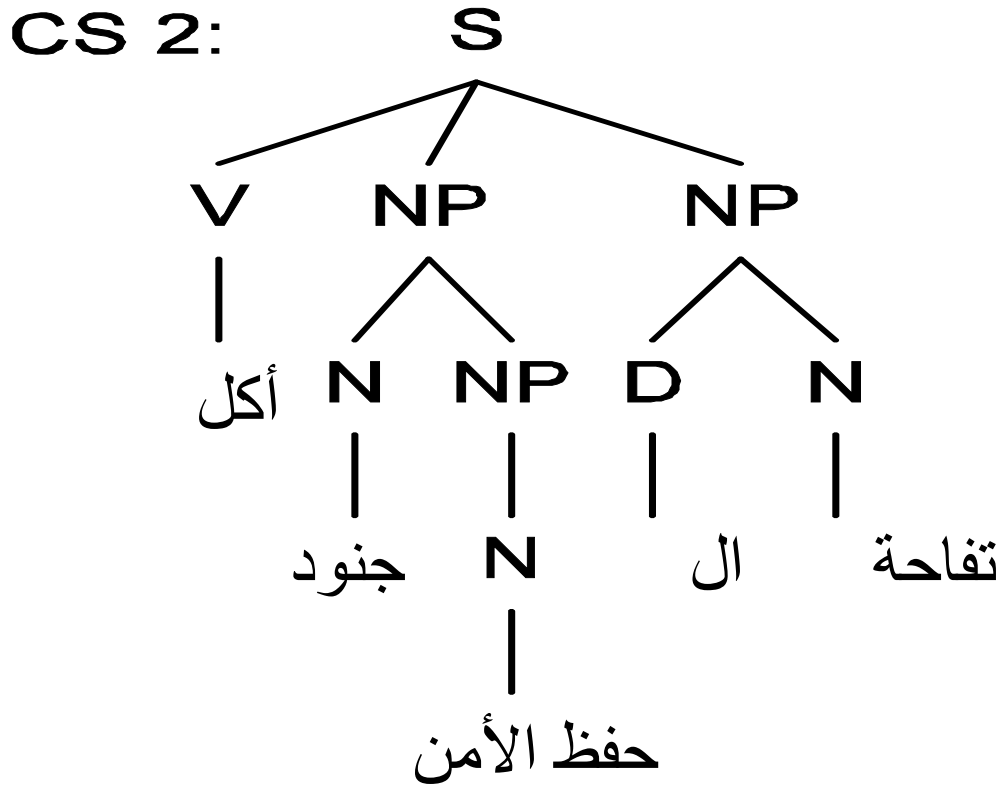
akal junud hifz al-amn al-tuffah

ate peace keeping soldiers the apples

(Peace keeping soldiers ate the apples)

Handling MWEs

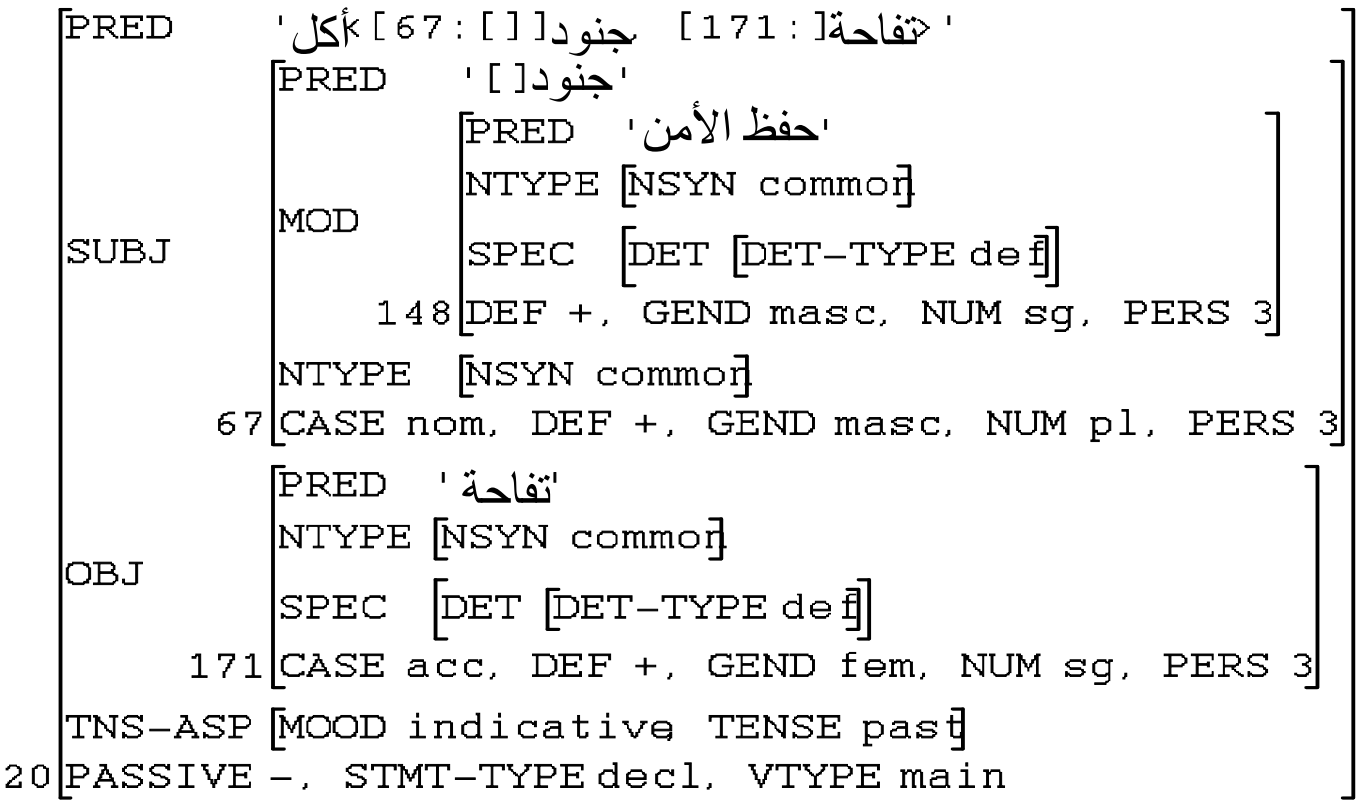
- C-Structure



Handling MWEs

- F-Structure

" "



Handling MWEs

- Syntactically-Flexible expressions

When a noun is modified by an adjective it usually allows for genitive nouns or pronouns to come in between

a دراجة نارية

darrajah nariyah

bike fiery

(motorbike)

b هذه دراجة الولد الصغير النارية

this darrajah al-walad al-saghir al-nariyah

this bike the-boy the-young the-fiery

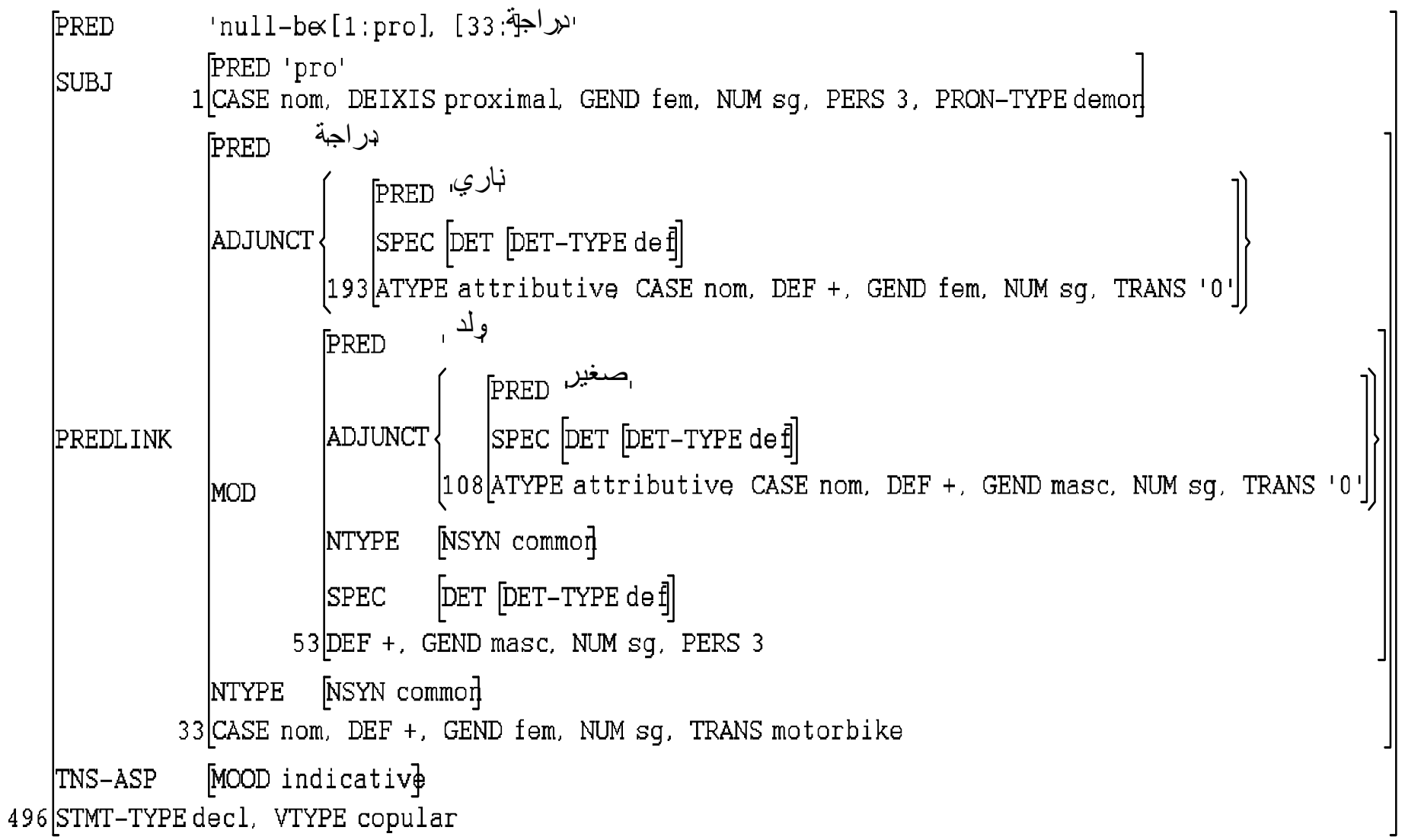
(This is the young boy's motorbike)

Handling MWEs

- Syntactically-Flexible expressions

```
bike   N XLE {
    (^PRED='bike' (^ ADJUNCT PRED)=c 'fiery'
                          (^ TRANS)=motorbike
    | (^PRED='bike' (^ ADJUNCT PRED)~= 'fiery'
                          (^ TRANS)=bike
    }.
```

"


PRED 'null-be<[1:pro], [33:دراجة]'
SUBJ
1
PRED 'pro'
CASE nom, DEIXIS proximal, GEND fem, NUM sg, PERS 3, PRON-TYPE demon
PREDLINK
PRED دراجة
ADJUNCT
PRED ناري
SPEC DET DET-TYPE def
193
ATYPE attributive, CASE nom, DEF +, GEND fem, NUM sg, TRANS '0'
MOD
PRED ولد
ADJUNCT
PRED صغير
SPEC DET DET-TYPE def
108
ATYPE attributive, CASE nom, DEF +, GEND masc, NUM sg, TRANS '0'
NTYPE NSYN common
SPEC DET DET-TYPE def
53
DEF +, GEND masc, NUM sg, PERS 3
NTYPE NSYN common
33
CASE nom, DEF +, GEND fem, NUM sg, TRANS motorbike
TNS-ASP MOOD indicative
496
STMT-TYPE decl, VTYPE copular

Handling MWEs

- Grammatically Flexible

Phrasal verbs in Arabic allow subjects to intervene between verbs and objects. This is why they need to be handled in the Syntax.

اعتمد الولد على البنت

i'tamada al-waladu 'ala al-bint

relied the-boy on the-girl

(The boy relied on the girl)

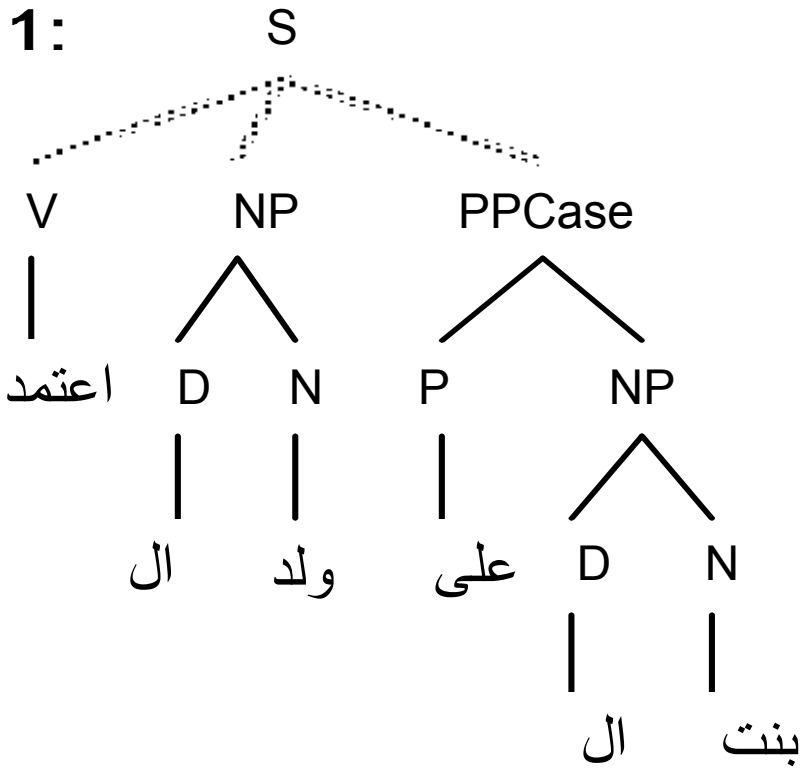
Handling MWEs

- Grammatically Flexible

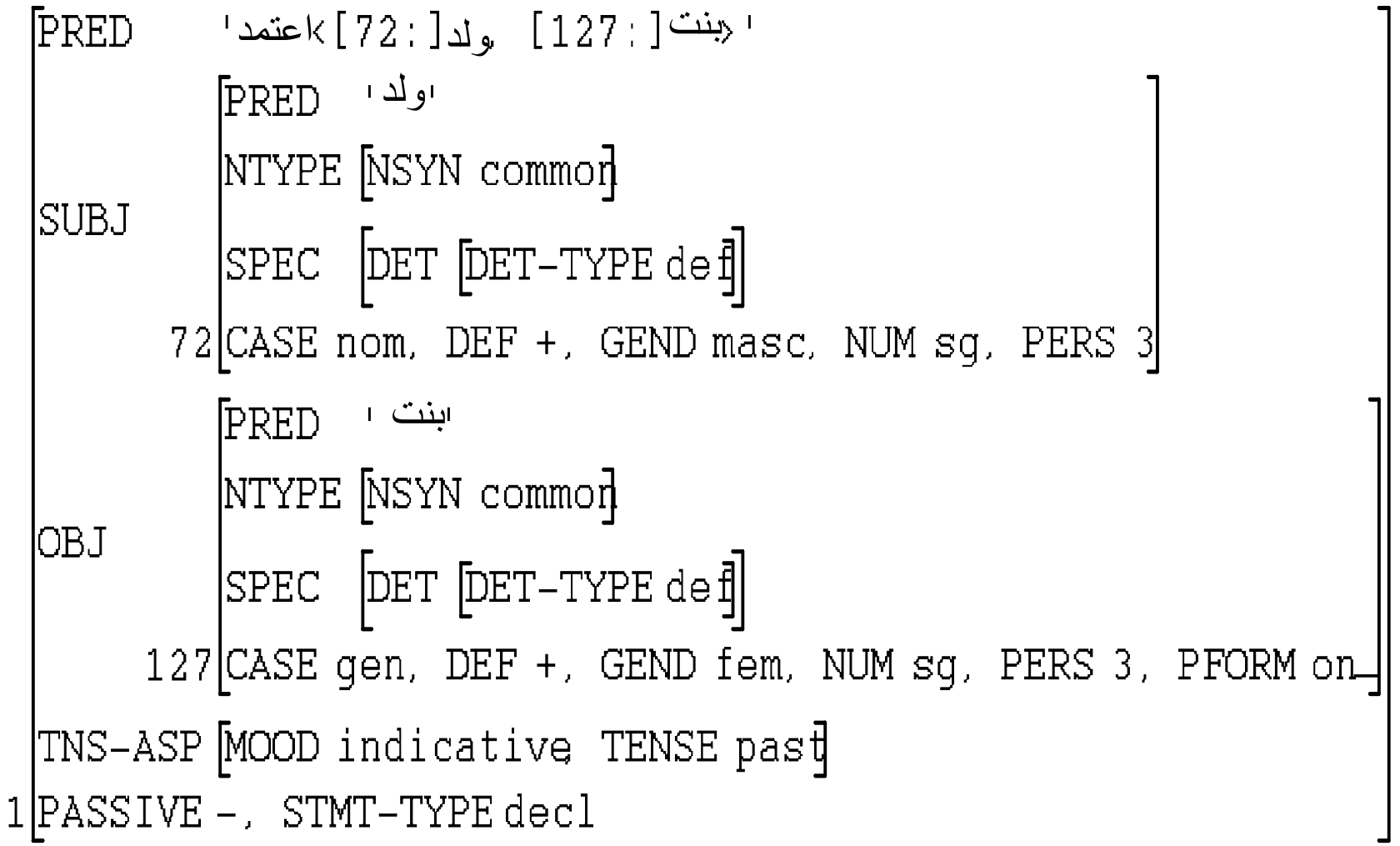
on P XLE (^ PFORM)=on (^ PCASE)=gen.

rely V XLE
(^ PRED)='rely<(^ SUBJ)(^ OBJ)>'
(^ OBJ PFORM)=c on.

CS 1:



" "


PRED 'اعتمد>[72:]ولد, [127:]بنت'
SUBJ
72
PRED 'ولد'
NTYPE [NSYN common]
SPEC [DET [DET-TYPE def]]
CASE nom, DEF +, GEND masc, NUM sg, PERS 3
OBJ
127
PRED 'بنت'
NTYPE [NSYN common]
SPEC [DET [DET-TYPE def]]
CASE gen, DEF +, GEND fem, NUM sg, PERS 3, PFORM on
TNS-ASP [MOOD indicative, TENSE past]
1
PASSIVE -, STMT-TYPE decl

Conclusion

- Normalizer: White spaces
- MWE Morphological Transducer
  - Tokenizer
  - Transduction
- Grammar: Lexical rules