

# Lexicon Development

## Case Study



Shafiq Ur Rahman

Center for Research in Urdu Language Processing  
National University of Computer and Emerging  
Sciences, Lahore

# Overview

- ▶ Computational Lexicon
- ▶ Tasks requiring Computational Lexicons
- ▶ Dimensions
- ▶ Lexicon Development Phases

# Lexicon

- ▶ It is the central repository of data for all language processing applications
- ▶ It contains information for human consumption as well as computer programs.

# Goal

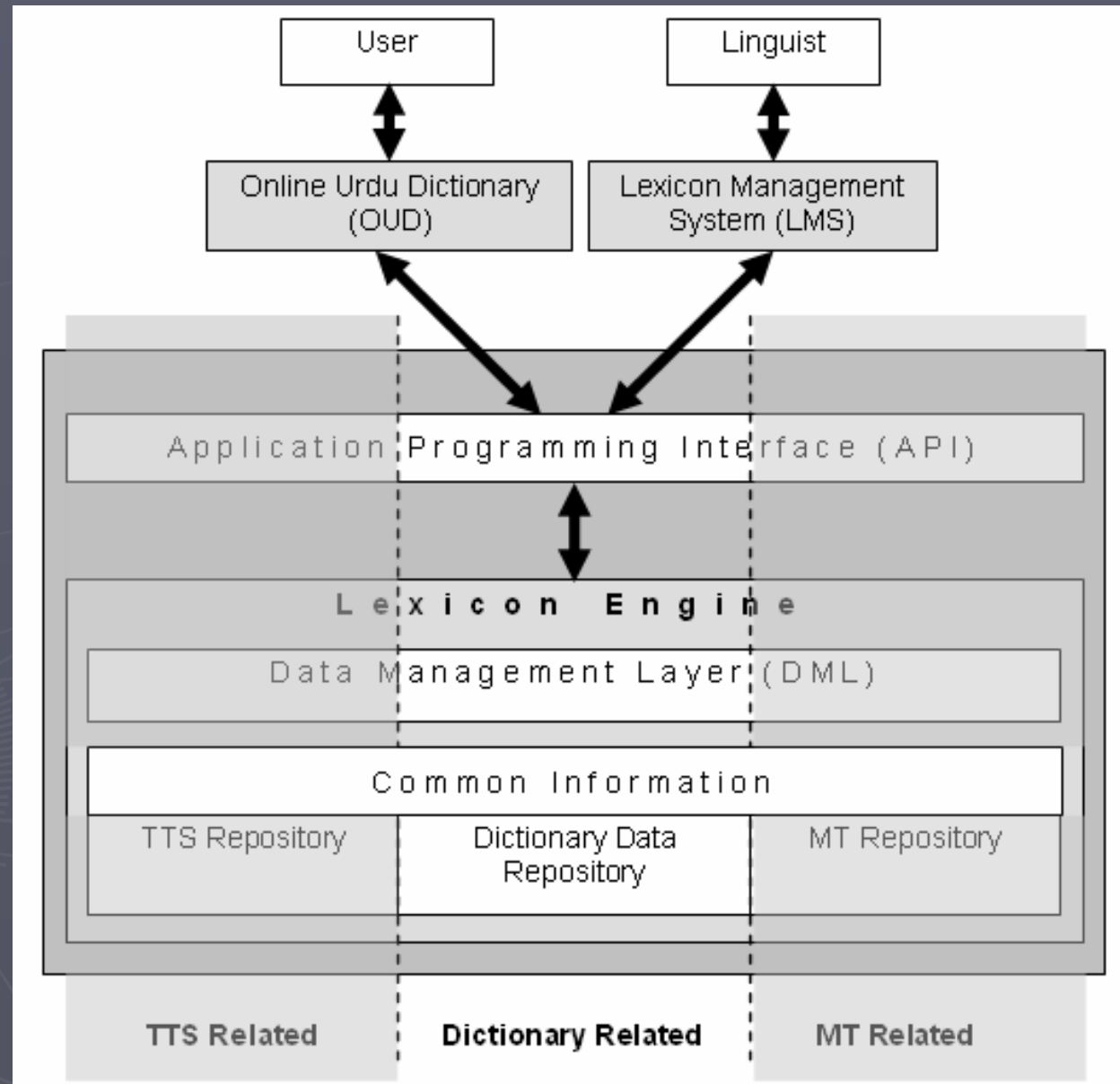
- ▶ An online Urdu dictionary of 50,000 words with complete contents for about 10,000 words
- ▶ Provide information for TTS
- ▶ Provide information for MT system

# Initial Decision

- ▶ General Purpose
- ▶ Application Specific



# Architecture of Lexicon



# Lexicon Entry Structure

- ▶ Do all words Qualify for being separate entries? For example
  - Inflections (boy → boys)
    - ▶ Gender based variations are inflections or derivations?

برا - بری

لڑکا - لڑکی

# Lexicon Entry Structure

## ▶ Pronunciations

- Format (diacritics, IPA)

- ▶ There is no standard way of marking diacritics in Urdu.

- should we provide pronunciation of inflections

# Lexicon Entry Structure

- ▶ What kind of relations between the entries should be maintained.
  - Morphological relations
    - ▶ Inflectional
    - ▶ Derivational
  - Semantic relations
    - ▶ Synonym
    - ▶ Antonym
    - ▶ Cross references

# Lexicon Entry Structure

- ▶ Words can have multiple POS
  - What kind of information is shared/not shared by different POS

# Lexicon Entry Structure

- Do all meanings of a word exist at same level?

1- بادل، گھٹا، بدلی۔

2- ہلکی نیلگوں چمکیلی دھوپ چھاؤں یا لہریں جو تلوار، خنجر وغیرہ کے پھل، ڈھال کی سطح، بندوق کی نال یا دوسرے فولادی اسلحہ پر صیقل سے پیدا ہوں، جوہر۔

3- چمکیلا لہریا جو رنگ یاروغن سے کپڑے یا کاغذ پر ڈالا جائے، جیسا کتاب کی جلد کی ابری میں ہوتا ہے۔

# Lexicon Entry Structure

## ► Compounds

- Should they make separate entries

خوبصورت

- Should they have any relation with their constituent words.

# Lexicon Entry Structure

## ► Idioms

“Nip the evil in the bud”

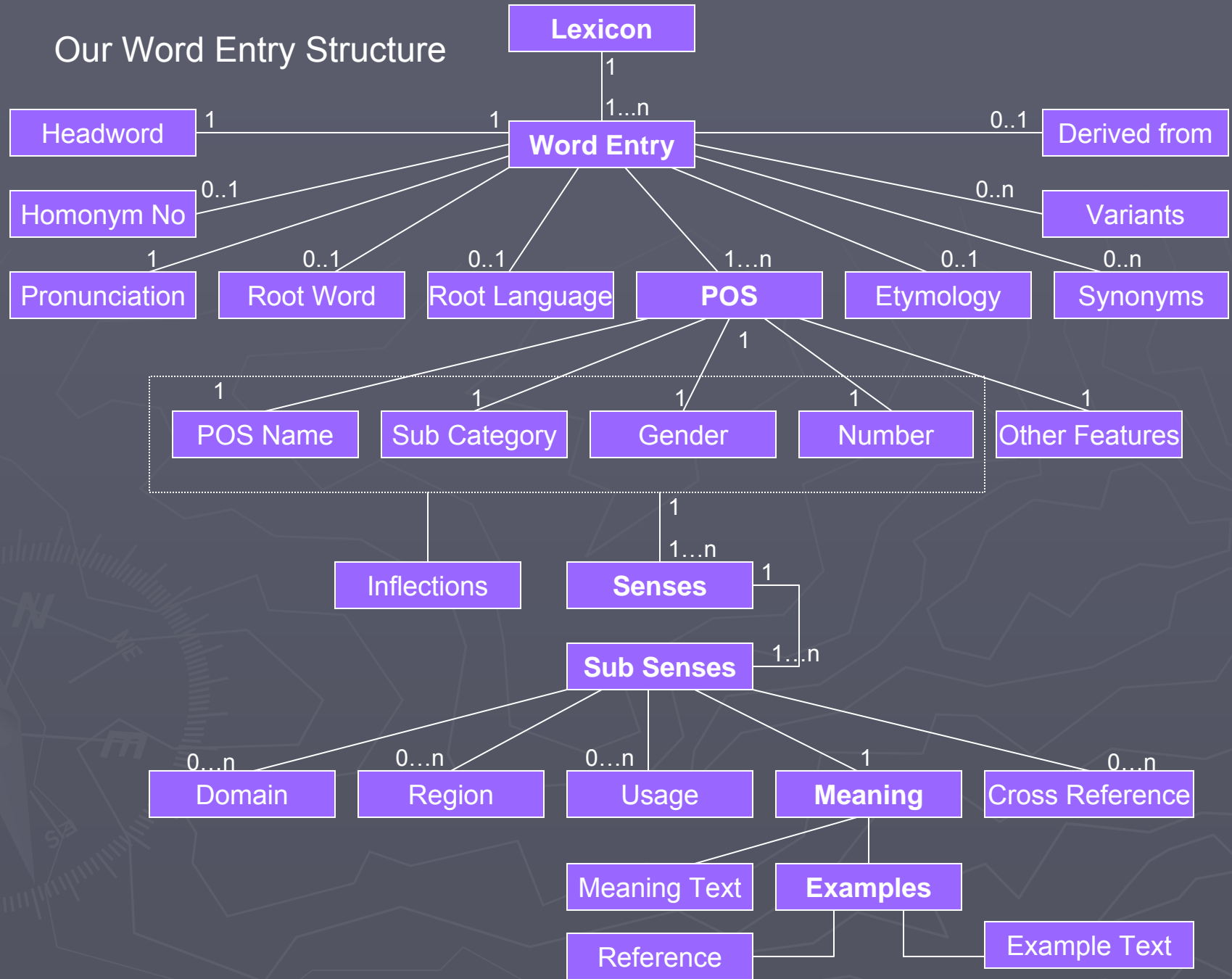
- Do they make separate entries
- Should they have any relation with words.

# Lexicon Entry Structure

## Information to be structured

- ▶ Headword
- ▶ Pronunciation
- ▶ Variant Forms
- ▶ Inflections
- ▶ Root Word
- ▶ Root Language
- ▶ POS Marker
- ▶ Sense
- ▶ Meaning
- ▶ Examples of Meaning
- ▶ Labels for Meaning
- ▶ Cross Reference
- ▶ Synonym/Antonym
- ▶ International Phonetic Alphabet (IPA)
- ▶ Etymology
- ▶ Idioms
- ▶ Derived From

# Our Word Entry Structure



# Data Gathering/Entry Process

## ▶ Data Gathering

- Manual
- Automatic (Corpus based)

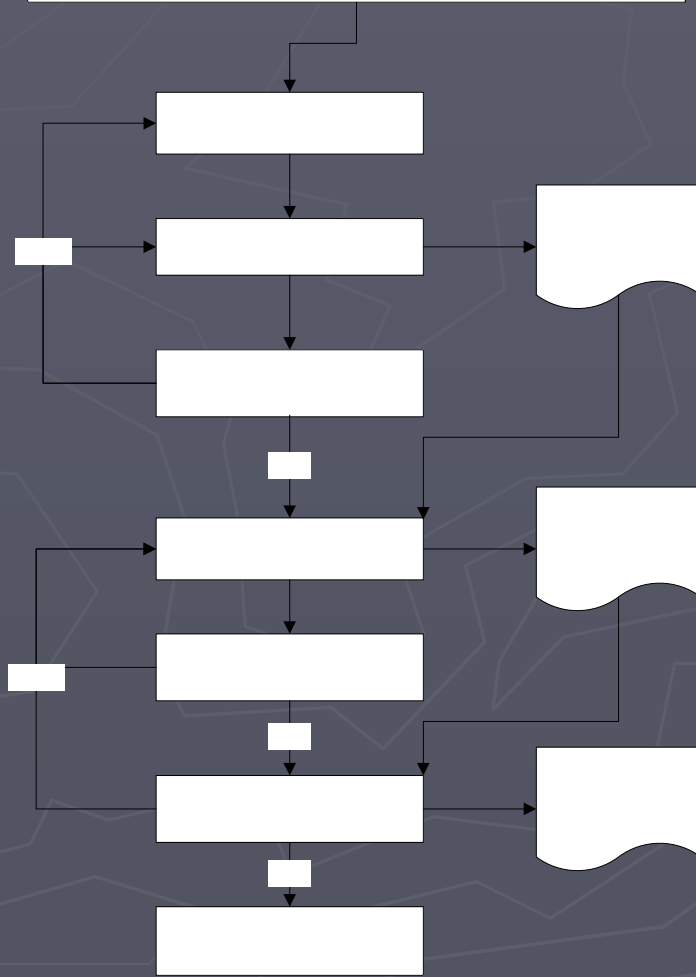
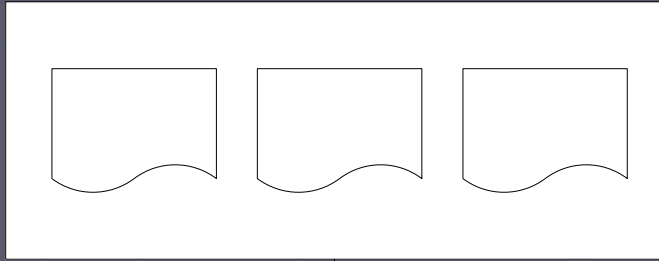
## ▶ Manual Data Gathering

- Tagged Data Entry Forms (Paper based)

# Data Gathering/Entry Process

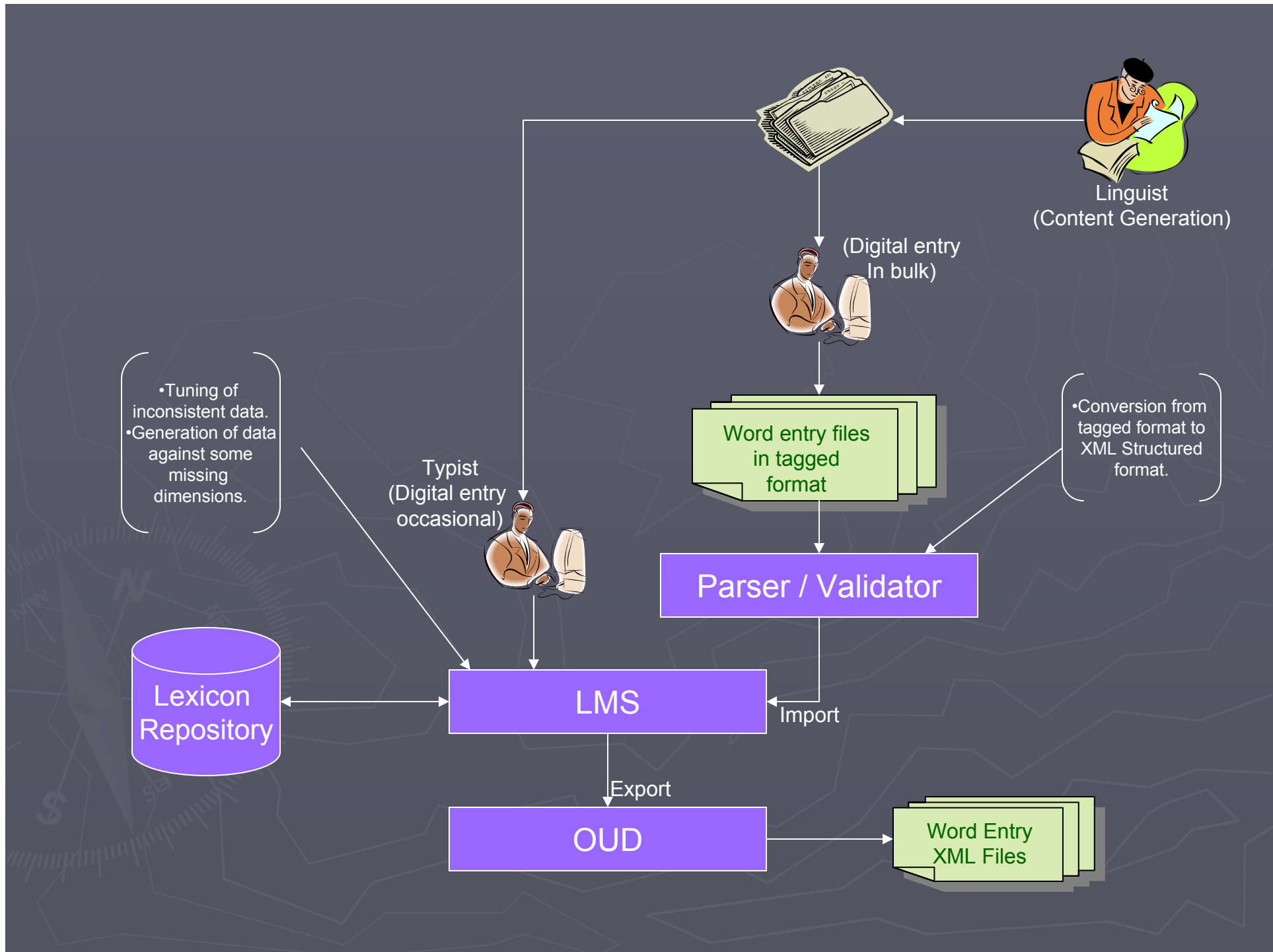
## ► Data Entry

- Only Tags and hand written data in data entry forms is typed in.
- Flat files (No Forms)



# Overall Process

- ▶ Words' Content Generation
- ▶ Digital Entry of Content
- ▶ Parsing of tagged word entry to form structured word entry (Parser)
- ▶ Processing on Structured word entries (Lexicon Management System)
- ▶ Transformation of word entries into the format required for Online Dictionary (Lexicon Management System).
- ▶ Presentation of dictionary data to online user. (Online Urdu Dictionary)



# Lexicon Management System

- ▶ Data Repository
  - RDBMS
- ▶ Database handling Layer
- ▶ API
- ▶ Interface

# Lexicon Management System

## ► Major Components

- Import
- Grinder
- Export
- Data Manipulation

شكریه

