Assignment 1

CNG463 - Introduction to Natural Language Processing
METU NCC Computer Engineering

Fall 2025-26

Submission Information

Due Date: 31 October 2025 (Friday) before midnight
Submission: Only through **ODTUClass** as a single PDF file
File Name: `initials_studentID_as1.pdf` (e.g., `gs_999999_as1.pdf`)
Content Requirements: Include your full name and student ID in the document
Length: 2-3 pages (no specific template required)
Critical:

- Email submissions will be deleted immediately and receive 0 points
- Late submissions are not permitted

Overview

This assignment explores two fundamental NLP tasks through hands-on experiments:

- Automated Translation
- Language Identification

Language Requirements: Use **English + one additional language** of your choice for all tasks.
Grading Philosophy: Each subtask has a baseline score for completing required examples and a creativity bonus for original work. Point distributions are specified below.

Task 1: Automated Translation (40 points)

Machine translation automatically converts text between languages. Modern neural systems handle syntax and semantics well but struggle with ambiguity, idioms, and cultural expressions.
Recommended Tools: DeepL, Google Translate, Bing Translator

Subtask 1.1: Basic Translation (10 points)

Baseline (7 pts): Translate at least 3 English sentences to your second language, then back to English. Use these examples:

- "This course is cool."
- "The pilot landed the plane."
- "I read the book on the table."

Identify: lexical ambiguities (e.g., "cool"), gender/role biases (e.g., "pilot"), and differences between forward and backward translations.

Creativity bonus (3 pts): Add 2 original sentences that test specific translation challenges.

Subtask 1.2: Back-Translation (10 points)

Baseline (7 pts): Translate "The spirit is willing but the flesh is weak" to your second language and back to English. Document whether meaning is preserved and what shifts occurred.

Creativity bonus (3 pts): Test 2 additional proverbs or culturally-specific expressions.

Subtask 1.3: Ambiguity and Idioms (10 points)

Baseline (7 pts): Test these English examples:

- Ambiguous: "Time flies like an arrow"
- Idioms: "It costs a fortune", "Break a leg", "Piece of cake"

Add 2 idioms from your second language. Determine whether translations are literal or idiomatic, and if meaning is preserved in round-trip translation.

Creativity bonus (3 pts): Test 2 additional ambiguous phrases or idioms that reveal system limitations.

Subtask 1.4: Creative Examples (10 points)

Baseline (7 pts): Create 2 original challenging sentences (e.g., cultural references, code-mixed text, figurative language, sarcasm).

Creativity bonus (3 pts): Add 2 more creative examples with insightful analysis of why they are challenging and how systems handle them.

Translation Key Points

- Errors occur at lexical, syntactic, semantic, and pragmatic levels
- Ambiguity and idioms challenge even advanced systems
- Creative test cases reveal limitations more effectively than standard examples

Task 2: Language Identification (40 points)

Language Identification determines the language of a text with high accuracy for clean inputs. Code-mixing, nonsense words, or short texts can reveal weaknesses.
Recommended Tools: LangId.py, fastText, Google Translate, DeepL

Subtask 2.1: Known Languages (5 points)

Baseline (5 pts): Test sentences in English and your second language. Verify correct identification.

Subtask 2.2: Mixed Sentences (10 points)

Baseline (7 pts): Create mixed-language sentences using both languages (e.g., "Bu software'de bug var ve crash yapiyor"). Replace content words with nonsense (e.g., "Bu xwt'de xwt var ve xwt yapiyor"). Analyse whether the system still detects the correct language and explain why.

Creativity bonus (3 pts): Test 2 additional mixed or modified sentences with insightful analysis.

Subtask 2.3: Grammar-Free Text (10 points)

Baseline (7 pts): Compare language detection for:

- "xwt xwt xwtyxwy xwt wxt" (no grammar words)
- "software bug crash sistem problem" (content words only)

Explain why results differ.

Creativity bonus (3 pts): Test 2 additional grammar-free examples with analysis of what features the system relies on.

Subtask 2.4: Fake Words (10 points)

Baseline (7 pts): Create made-up words that resemble English or your second language:

- Turkish-looking: "programlalik", "yazilimgil", "sisteminlerci"
- English-looking: "computerish", "teknomagy"

Analyse how the system classifies them and what features it relies on.

Creativity bonus (3 pts): Create 2 additional fake words with analysis of classification patterns.

Subtask 2.5: Creative Examples (5 points)

Baseline (3 pts): Design 1 original test that challenges language identification systems.

Creativity bonus (2 pts): Add 1 more creative example with insightful analysis.

How Language Identification Works

- **Special Characters:** (e.g., "s with cedilla", "g with breve" in Turkish)
- **Frequent Words:** (e.g., "the", "bir", "ve")
- **Character N-grams:** statistical patterns unique to each language

Even nonsense words may resemble a language through character patterns.

Task 3: Reflection and Insights (20 points)

Baseline (14 pts): Provide a thoughtful reflection addressing:

- Key surprises or unexpected behaviours
- Most difficult cases for systems
- Lessons learned about language and NLP

Creativity bonus (6 pts): Demonstrate deeper analysis through:

- Comparative system analysis (if multiple tools used)
- Connections to linguistic concepts
- Practical implications for real-world applications

Report Structure

Your 2-3 page report should include:

1. **Translation Experiments:** Systems used, sentences tested, error types observed, analysis
2. **Language Identification Experiments:** Systems used, observations for different input types, explanations based on system features
3. **Reflection:** Key insights, challenging cases, and lessons learned

Grading Summary

Total: 100 points

- Task 1 (Translation): 40 pts (28 baseline + 12 creativity)
- Task 2 (Language ID): 40 pts (29 baseline + 11 creativity)
- Task 3 (Reflection): 20 pts (14 baseline + 6 creativity)

Note: Completing only required examples yields a maximum 71/100. The remaining 29 points reward originality, deeper analysis, and creative exploration.