

# Exploratory Data Analysis Report

This report synthesizes the findings from the Exploratory Data Analysis (EDA) of 5 distinct Entity Linking (EL) datasets: **AIDA-EE** (**apw\_eng\_201010.tsv & apw\_eng\_201011.tsv**), **KORE50**, **N3 (RSS-500 & Reuters-128)**, **MSNBCt**, and **OKE Challenge**. The analysis reveals significant diversity in dataset size, annotation format, density, and task complexity, highlighting their different potential applications for training and evaluating EL models.

---

## 1. AIDA-EE Dataset Analysis

The AIDA-EE dataset consists of 298 documents with 9,976 total annotations, making it a large and densely annotated corpus.

- **Format:** Span-based annotations in a TSV format, linking mention strings to YAGO2 IDs and Wikipedia URLs.
  - **Density:** The dataset is very dense, with an **average of 33.5 mentions per document** (ranging from 1 to 242).
  - **NIL Entities:** A key feature is the explicit tracking of "Emerging Entities" (NILs), which are tagged as `--00KBE--`. These account for **5.74% of all mentions** (573 total).
  - **Mentions:** The average annotated name length is 8.6 characters. The most common mention strings are abbreviations and major locations, such as `U.S.` (216 mentions), `China` (146), and `Japan` (104).
- 

## 2. KORE50 Dataset Analysis

The KORE50 dataset is a small, specialized dataset containing 50 documents.

- **Format:** This dataset is **token-based**, not span-based. It contains 160 total annotated tokens.
  - **Annotations:** It uses a "B/I" (Begin/Inside) tagging scheme. The analysis shows 148 "B" tokens and 12 "I" tokens, which constitute 132 unique full mentions.
  - **NIL Entities:** It tracks NIL entities using the `--NME--` tag. This dataset has a very low NIL rate of **3.12%** (5 total tokens).
  - **KB & Mentions:** It links to YAGO2. Mentions are short, with an average length of 6.3 characters. Top-linked entities at the token level include `Bob_Dylan`, `Stanford_University`, and `Steve_Jobs` (3 tokens each).
- 

## 3. N3 Datasets Analysis (RSS-500 & Reuters-128)

This analysis covered two distinct NIF-based datasets (RSS-500 and Reuters-128) that both link to DBpedia URIs. Their structures, however, are vastly different.

### RSS-500

- **Size:** 500 short text fragments.
- **Density:** This dataset has a **fixed, artificial density**. Every single fragment contains **exactly 2 tagged entities**.
- **Text:** Fragments are short, averaging 31 words.
- **Entities:** Contains 849 unique entities. The most common is [Associated\\_Press](#).

## Reuters-128

- **Size:** 128 long text fragments.
  - **Density:** This dataset is much more variable and dense, with an **average of 6.9 entities per fragment** and a maximum of 43.
  - **Text:** Fragments are significantly longer, averaging 124 words.
  - **Entities:** Contains 444 unique entities. Top entities include [Dominion\\_Textile](#) and [Japan](#).
- 

## 4. MSNBCt Dataset Analysis

The MSNBCt dataset is a large but sparse NIF-based corpus.

- **Size:** It contains 1,231 text fragments, the largest number of documents in this analysis.
  - **Density:** It is **extremely sparse**, with an average of only **0.6 entities per fragment**. The histogram shows that a vast number of fragments contain 0 entities.
  - **Text:** Fragments are short, averaging 39.5 words.
  - **KB & Entities:** It links to Wikipedia URLs. It contains 328 unique entities, with [Gerald\\_Ford](#) and [Nick\\_Saban](#) being the most common (25 mentions each).
- 

## 5. OKE Challenge Dataset Analysis

The OKE dataset is a medium-sized NIF corpus with 101 sentences (contexts) and 664 total annotations. It presents the most complex task structure.

- **Key Feature 1: Entity Typing:** This is the only dataset that explicitly classifies its 373 unique entities into types:
  - **DUL:Person:** 165 entities
  - **DUL:Organization:** 119 entities
  - **DUL:Place:** 89 entities
- **Key Feature 2: Co-reference:** The dataset includes annotations for **pronouns**. The most frequent "anchors" are [he](#) (31) and [his](#) (31), indicating a task that blends EL with co-reference resolution.
- **NIL Entities:** The dataset has a high rate of NIL entities. Of the 373 unique entities, 97 are missing DBpedia links, resulting in a **NIL rate of 26%**.
- **Density:** The dataset is fairly dense, with an average of **6.6 annotations per sentence**.