Named Entity Recognition in Football Journalism: A Comparative Analysis of Men's and Women's Coverage

Ziv

February 2026

Abstract

This report presents a comprehensive Named Entity Recognition (NER) pipeline applied to football journalism, with the goal of testing fifteen hypotheses about entity prominence, club co-occurrence, and gendered framing differences between men's and women's football coverage. Our data pipeline combines web scraping from four independent sources (GDELT, BBC RSS, direct site crawling, and dedicated women's football outlets) with seven Kaggle football datasets, yielding a combined corpus of over 20,000 unique articles. After cleaning, deduplication, football-relevance filtering, and gender classification, we retain 11,386 men's and 563 women's articles from Kaggle, supplemented by 2,316 scraped women's articles to address class imbalance. We evaluate three spaCy NER models on the SocCor UEFA EURO 2024 corpus (121 documents, 13,353 gold player mentions) and select `en_core_web_lg` (F1 = 0.812) as our primary model. Detailed error analysis reveals that 40.4% of shared model blind spots involve single-word surname references, and players from Eastern/Central European countries are disproportionately missed. Through gazetteer-enhanced entity classification and context-window hypothesis testing with formal statistical tests (Mann-Whitney U, Chi-square, proportion z-tests) and effect-size reporting (Cohen's d, Cramér's V, Cohen's h), we find strong evidence for gendered differences in naming conventions (H10), meta-discourse about league growth (H14), and attribute framing (H12), alongside classical NER results confirming Zipfian player prominence (H1) and transfer-window MONEY spikes (H4). All p-values are corrected for multiple comparisons using Benjamini-Hochberg FDR adjustment.

Contents

1 Introduction

Named Entity Recognition is a fundamental NLP task that identifies real-world referents—persons, organisations, locations—in unstructured text. In the football domain, accurate NER enables downstream analyses such as player prominence tracking, club co-occurrence network construction, and gendered framing analysis.

This project addresses two overarching research questions:

RQ1. What structural patterns emerge from entity distributions in football journalism (e.g., player prominence, club clustering, transfer windows)?

RQ2. How does the *framing* of entities differ between men's and women's football coverage in terms of naming conventions, descriptive language, and meta-discourse?

We operationalise these questions through fifteen testable hypotheses grouped into **core football** (H1–H5) and **men vs. women portrayal** (H9–H16), plus a diagnostic analysis (D1) on entity diversity.

2 Data Sources

Our data comes from two broad categories: (1) pre-existing Kaggle datasets covering general football journalism using which we run the hypothesis analysis, and (2) a purpose-built web-scraping pipeline targeting Champions League and women's football coverage. Additionally, we use the SocCor corpus as ground truth for NER evaluation.

2.1 Source 1: Kaggle Football Datasets

We aggregated seven CSV files from the `data/kaggle_data_football/` directory, each with a different schema (Table 1).

Table 1: Kaggle data sources and article counts.

File	Raw rows	Primary source	Coverage
`final-articles.csv`	11,963	Goal.com, multi-source	General football
`goal-news.csv`	7,596	Goal.com	News, match reports
`df_analyst.csv`	1,960	The Analyst	Analytics, reviews
`tribuna_articles.csv`	1,500	Tribuna.com	Fan community
`allfootball.csv`	800	All Football App	Mixed
`skysports.csv`	278	Sky Sports	Premier League focus
`live_mint.csv`	40	LiveMint	Indian perspective
Total (raw)	24,137		

2.2 Source 2: Web Scraping Pipeline

To supplement the Kaggle data—particularly for women's football, which is under-represented in general datasets—we built a four-stage scraping pipeline, orchestrated by `run_all_scrapers.py` and configured centrally via `config.py`.

2.2.1 Stage A: GDELT API Scraping (`1_scrape_gdelt.py`)

The GDELT DOC 2.0 API was queried using sliding 60-day date windows from January 2023 to February 2026. For **men's coverage**, 15 queries targeted Champions League content (e.g., `"Champions League" match report football`, `"Champions League" football transfer`). For **women's coverage**, 37 queries spanned UWCL, WSL, Liga F, NWSL, Women's World Cup, and individual club teams (e.g., `"Women's Champions League" football`, `"WSL" football match report`, `"Barcelona Femeni" football`, `"Lionesses" football England women`).

Each GDELT query returns up to 250 article URLs. Full article text was then extracted using Trafilatura with concurrent thread-pool downloading (6 workers), respecting rate limits of 2 s between GDELT queries and 1 s between Trafilatura fetches. Blocked domains (e.g., Irish regional newspapers returning non-football content) were filtered out. Articles were saved as individual JSON files with metadata including URL, title, date, domain, source country, language, and word count.

Result: 11,547 men's and 14,712 women's raw JSON articles.

2.2.2 Stage B: BBC Sport RSS (`2_scrape_bbc_rss.py`)

BBC Sport RSS feeds were parsed for Champions League, European Championship (men's), and Women's Football sections. For each entry, Trafilatura extracted the full article body. Articles shorter than 50 words were discarded. Duplicate URLs were tracked across runs to avoid re-downloading.

Result: 65 men's and 34 women's articles.

2.2.3 Stage C: Direct Website Crawling (`3_scrape_web_direct.py`)

Seed URLs from UEFA.com (UCL and UWCL news pages) and Goal.com (Champions League and women's football sections) were crawled to discover article links matching domain-specific URL patterns. Discovered URLs were then processed through Trafilatura for full-text extraction.

Result: 11 men's and 14 women's articles.

2.2.4 Stage D: Expanded Women's Scraping (`6_scrape_women_expanded.py`)

Recognising the persistent class imbalance, an expanded women's football scraper targeted additional sources: BBC Sport Women's Football (sitemap/archive discovery), SkySports Women's Football news pages, re-queried GDELT with simpler terms to catch previously missed articles, and Trafilatura's sitemap discovery for known women's football domains. This scraper maintained a shared URL deduplication set across all women's source directories.

2.3 Source 3: SocCor UEFA EURO 2024 Corpus

The **SocCor** corpus is a multilingual football commentary corpus annotated with inline player tags of the form `<SURNAME_POSITION_COUNTRY>` (e.g., `<BELLINGHAM_ATTACKING-MIDFIELD_ENG>`). We use the **English BBC** subset (121 documents, 13,353 gold player mentions) as ground truth for NER model evaluation. The corpus spans match reports, live game commentaries, highlights, and livetickers from the UEFA EURO 2024 tournament.

SocCor also provides metadata:

- `Players.csv`: 646 players with full name, token mapping, position, club, and market value.
- `Groups.csv`: Participating national teams grouped by tournament stage.
- `Broadcasters.csv`: Media outlets per country.

3 Data Processing Pipeline

The processing pipeline consists of seven stages, implemented as standalone scripts.

3.1 Stage 1: Scraped Data Cleaning and Merging

Script: `5_clean_and_merge.py`

All raw scraped JSON files from every source subdirectory (`gdelt/`, `bbc/`, `web_scraped/`) were loaded and merged per gender category. The cleaning pipeline applied three text transformations in sequence:

1. **HTML stripping:** Residual HTML tags were removed using BeautifulSoup with a regex fallback.
2. **Encoding repair:** Common mojibake artefacts (e.g., `\u00e2\u0080\u0099` → apostrophe, non-breaking spaces → regular spaces) were corrected, followed by Unicode NFC normalisation.
3. **Whitespace normalisation:** Multiple consecutive whitespace characters were collapsed into single spaces.

Articles with fewer than 50 words were discarded. Three-stage deduplication then removed duplicates by: (1) exact URL match, (2) case-insensitive title match, and (3) near-duplicate body text (first 200 characters match).

Result: 6,585 men's and 2,316 women's cleaned, deduplicated scraped articles, saved to `data/final_processed/`.

3.2 Stage 2: Kaggle Schema Normalisation and Deduplication

Script: `classify_kaggle.py`

Each of the seven Kaggle CSVs has a different column schema. Per-file column mappings normalised all data into a unified schema: `article_id | title | body_text | link | author | publish_time | source`. Missing columns were filled with empty strings. Deduplication on `(title, link)` removed 11,970 duplicate rows (largely overlapping articles between `final-articles.csv` and `goal-news.csv`). Articles with fewer than 50 characters of body text were discarded.

Result: 24,137 raw → 12,167 deduplicated → **11,949 final Kaggle articles**.

3.3 Stage 3: Gender Classification

Since no ground-truth gender labels exist in the Kaggle data, we applied **keyword-based classification** using a compiled regular expression with 40+ patterns:

- **Direct indicators:** `women's football`, `WSL`, `NWSL`, `UWCL`, `Liga F`, etc.
- **Team names:** `Lionesses`, `USWNT`, `Matildas`, `Banyana Banyana`, etc.
- **Club variants:** `Chelsea Women`, `Barcelona Femení`, `Arsenal Women`, etc.
- **Competition names:** `She Believes Cup`, `Arnold Clark Cup`, `Women's World Cup`, etc.

Both `title` and `body_text` were searched. An article was labelled *women* if any pattern matched; otherwise *other* (predominantly men's).

Result: Women's: 563 articles (4.7%), **Other (men's):** 11,386 articles (95.3%).

3.4 Stage 4: Final Input Assembly

Script: `prepare_final_inputs.py`

Given the severe class imbalance in the Kaggle data (563 women's vs. 11,386 men's), we combined the Kaggle women's articles with the cleaned scraped women's articles (§3.1) to create a larger women's corpus. Schema normalisation mapped the scraped columns (`url`, `domain`, `date`) to the Kaggle schema (`link`, `source`, `publish_time`). Title-based deduplication between the two sources ensured no overlap.

For men's articles, **only the Kaggle data** was used in the final analysis. While the scraping pipeline successfully collected 6,585 men's articles, computational resource con-

straints (NER processing time and memory) prevented us from running the full pipeline on the combined corpus. The scraped men's data is available for future analysis.

Final corpus for NER analysis:

- **Men's/Other:** 11,386 articles (Kaggle only)
- **Women's:** Kaggle (563) + scraped (2,316) = combined corpus (after deduplication)

3.5 Stage 5: NER Model Evaluation on SocCor

Script: `8b_evaluate_soccor.py`

Before running NER on our corpus, we evaluated model accuracy using SocCor as ground truth. Figure 1 illustrates the evaluation pipeline.

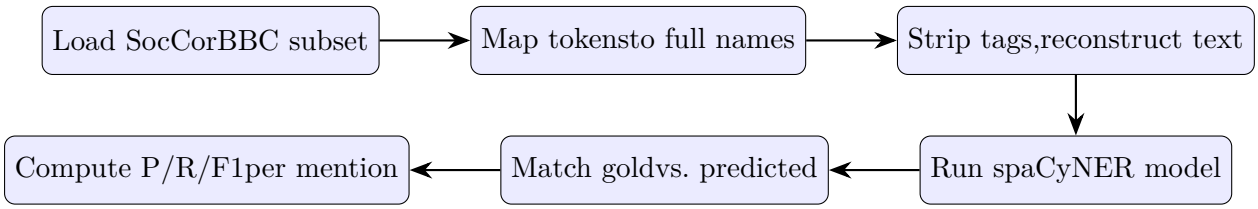

Figure 1: SocCor NER evaluation pipeline. Inline tags are replaced with full player names from `Players.csv`, then spaCy predictions are matched against gold entities using exact name, surname, or substring containment.

The pipeline operates as follows:

1. **Token-to-name mapping:** `Players.csv` (646 entries) maps inline tags to full player names (e.g., `<STEGEN_GOALKEEPER_GER>` → "Marc-André ter Stegen" instead of the fallback "Stegen").
2. **Text reconstruction:** Inline tags are stripped and replaced with the mapped full name, producing clean text suitable for NER.
3. **NER execution:** Each spaCy model processes the clean text, extracting all `PERSON`-labelled entities.
4. **Matching:** For each gold entity, we check whether any spaCy `PERSON` entity matches by: (a) exact full name, (b) mutual substring containment, (c) surname match, or (d) individual name-part match.
5. **Metrics:** Precision (*what fraction of spaCy PERSON predictions are real players*), recall (*what fraction of gold players were detected*), and F1 are computed at the mention level.

Table 2: SocCor NER evaluation results (121 BBC documents, 13,353 gold player mentions).

Model	Precision	Recall	F1	TP	FN	FP
`en_core_web_sm`	0.749	0.758	0.754	10,127	3,226	3,389
`en_core_web_lg`	0.754	**0.880**	**0.812**	11,749	1,604	3,832

Decision: `en_core_web_lg` was selected as the primary model, offering +7.6% F1 and +12.2% recall over `en_core_web_sm`. Its substantially higher recall (0.880 vs. 0.758) was particularly important for our downstream analysis, which relies on detecting as many player mentions as possible.

3.6 Stage 5a: SocCor Error Analysis

Script: `8c_soccor_error_analysis.py`

To understand *where* and *why* the NER models fail, we conducted a detailed error analysis that recorded every individual false negative (missed player) and false positive (spurious detection) with full context.

Most-missed players. Table 3 shows the players most frequently missed by each model. Mbappé is the single most-missed player (262 times by `sm`, 209 by `lg`), primarily because his name appears in live commentary as a bare surname—a pattern that spaCy's general English models struggle with for non-English names.

Table 3: Top-5 most frequently missed players by each model.

Player	Country	Position	sm misses	lg misses
Mbappé	FRA	Centre Forward	262	209
Hernandez	FRA	Left Back	132	—
Dembélé	FRA	Right Winger	119	—
Denzel Dumfries	NED	Right Midfield	102	102
Koundé	FRA	Centre Back	77	42
Memphis Depay	NED	Centre Forward	42	63
Lamine Yamal	ESP	Right Winger	—	57

Shared blind spots. 451 mention-level instances were missed by *both* models. Pattern analysis revealed:

- **40.4% are single-word surnames** (e.g., "Mbappé", "Kanté", "Koundé"): spaCy struggles with bare surnames lacking a preceding first name, especially for non-English names with diacritics.
- **27.9% are Eastern/Central European players** (Turkey, Romania, Georgia, Slovakia, etc.): names with unfamiliar morphology for English-trained models (e.g., Barış Alper Yılmaz, Milan Škriniar).
- **Centre backs are disproportionately missed** (121 of 451 shared misses): defenders' surnames are commonly used without first names in rapid live commentary.

Document type matters. Games/live commentary text accounts for **93.6%** of all false negatives (3,019/3,226 for `sm`; 1,466/1,604 for `lg`). This is because live commentary

uses informal, rapid-fire surname-only references ("Mbappé drives forward") far more than match reports, which tend to introduce players by full name.

Common false positives. The most frequent spurious `PERSON` detections included: "Euros" (96 times by `sm`—a tournament name misclassified), "Pepe" and "Mudrik" (real players but not in the gold annotation set for the specific document), "Austria" (country name misclassified), "Goal" (common noun), and referee names like "Anthony Taylor" and "Michael Oliver" (correctly `PERSON` but not in the gold *player* set).

3.7 Stage 6: NER Extraction with Gazetteer Enhancement

Script: `7_run_ner.py`

The NER pipeline processes each article through a carefully designed classification cascade:

1. **Football relevance filtering:** Each article is checked against 100+ domain keywords (e.g., "goal", "transfer", "league", "penalty", "offside"). Articles matching fewer than 3 keywords are classified as non-football and skipped. Additionally, articles containing American sports terms ("touchdown", "quarterback", "inning") are rejected.
2. **SpaCy NER:** `en_core_web_lg` extracts all named entities from the article body (truncated to 10,000 characters for memory safety). The `tagger`, `parser`, and `lemmatizer` components are disabled for speed, processing articles in batches of 100 via `nlp.pipe()`.
3. **Domain-specific reclassification:** Each spaCy entity passes through a priority cascade:
   a) **Noise filter:** Non-football entities ("Trump", "NFL", "Tesla", etc.—a curated list of 35+ terms) are removed entirely.
   b) **Gazetteer matching:** Entity text is normalised (lowercase, quote stripping, article removal, whitespace collapse) and checked against five separate gazetteers in priority order:
      - Competition (40+ entries): "Champions League", "WSL", "Copa del Rey", etc.
      - Media (35+ entries): "Sky Sports", "BBC Sport", "ESPN", etc.
      - Club (130+ static entries + 219 from SocCor `Players.csv` club column): "Chelsea", "Barcelona", "Portland Thorns", etc.
      - Governing Body (20+ entries): "FIFA", "UEFA", "The FA", etc.
      - Player (1,279 entries from SocCor `Players.csv`—full names + surnames): "Florian Wirtz", "Wirtz", "Mbappé", etc.
   c) **SpaCy-label fallback:** Unmatched entities use their spaCy label (e.g., `PERSON` → Player,

ORG → Org_Other, GPE → Location, MONEY → Money).

4. **Context capture:** For each entity, a ±50-character window around the mention is saved as `sentence_context`, enabling downstream hypothesis testing on the language *surrounding* entity mentions.

Key design decision: The Player gazetteer (from `Players.csv`) is kept **completely separate** from the Club gazetteer. Player names (1,279 normalised entries) and club names (130+ static + 219 from the club column) never overlap, preventing misclassification of players with club-like names.

Result:

- **Other (men's):** 37,089 entities from 718 articles (9,652 unique); 282 articles filtered as non-football
- **Women's:** 24,464 entities from 369 articles (7,050 unique); 194 articles filtered as non-football

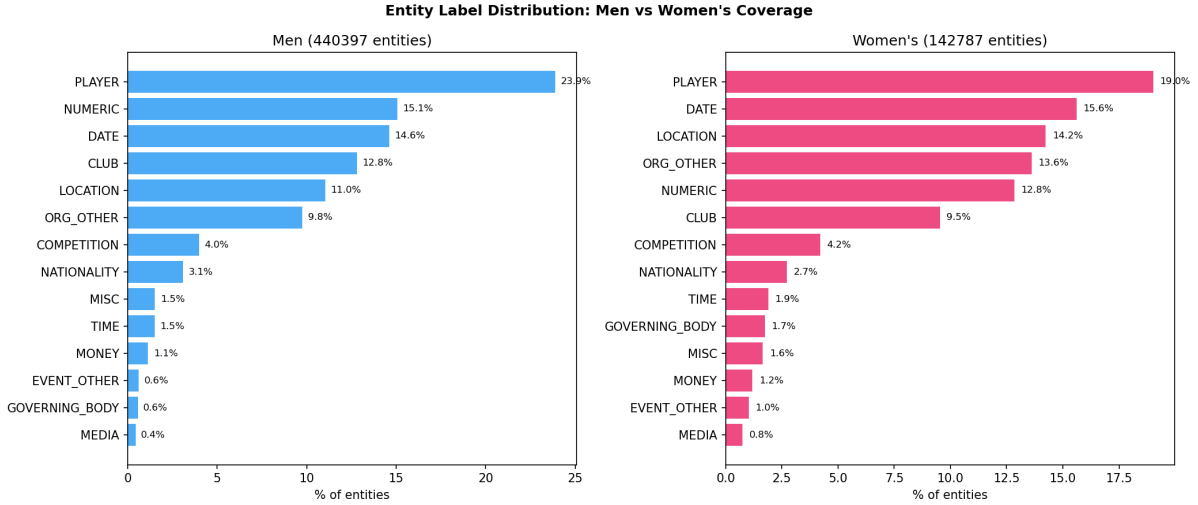

Figure 2: Entity label distribution comparison between men's and women's coverage. The Player label dominates both, but men's shows a higher proportion of Player mentions relative to other entity types.

4 Hypotheses

4.1 A. Core Football Entity-Occurrence Hypotheses

These hypotheses examine structural patterns in entity distributions, testable via NER counts and co-occurrence analysis.

H1. Prominence. A small number of Player entities dominate mentions within each category, following a heavy-tail Zipfian distribution (e.g., the top-10 players account for a disproportionate share of mentions).

H2. **Club centrality.** CLUB co-occurrence networks form distinct clusters by competition (domestic league vs. continental tournament), and the purity of these clusters differs between men's and women's coverage.

H3. **Manager-focus.** In articles with manager/coach mentions, negative event terms ("sacked", "under pressure", "crisis") co-occur more frequently, and the rate of such co-occurrence differs between men's and women's coverage.

H4. **Transfer windows.** In transfer-related articles, PLAYER $\leftrightarrow$ CLUB co-occurrences increase sharply and MONEY mentions spike, with measurable effect sizes.

H5. **Injury narrative.** Articles exhibit a rich injury/fatigue vocabulary; women's coverage may show different injury patterns (particularly ACL-related terms) compared to men's.

4.2 B. Men vs. Women Portrayal Hypotheses

These hypotheses examine *how* entities are framed differently, testable via context-window analysis around entity mentions.

H9. **Individual vs. team emphasis.** Women's reports contain fewer distinct PLAYER entities per article and a different player-to-club ratio compared to men's.

H10. **Name formality.** Women's articles use full names (first + last) for PLAYER at a higher rate; men's articles use surname-only at a higher rate.

H11. **Role framing.** Women's articles mention youth/age descriptors near PLAYER more often; men's mention experience/legacy terms more often.

H12. **Attribute mix.** Women's coverage uses more mentality/effort descriptors ("brave", "determined") near PLAYER; men's uses more physicality/tactical descriptors ("powerful", "clinical").

H13. **Relational framing.** Women's coverage includes more relational mentions near PLAYER ("captain", "teammate") than men's, which favours "star" framing.

H14. **Meta-discourse.** Women's coverage contains more "league growth / visibility / professionalisation" terms co-occurring with COMPETITION/CLUB.

H15. **Credit assignment.** In win-related articles, men's reports attribute success to individual PLAYER entities more often; women's reports credit the collective.

H16. **Source framing bias.** Different news outlets (Goal.com, The Analyst, Tribuna, etc.) exhibit different entity framing patterns, measurable via Kruskal-Wallis tests across sources.

4.3 C. Diagnostic Analysis

D1. Entity diversity. Shannon entropy of entity label distributions per article differs between men's and women's coverage, reflecting differences in coverage breadth.

5 Methodology

All hypothesis testing is implemented in `10_hypothesis_analysis.py`. The general approach combines context-window linguistic analysis with formal statistical testing.

5.1 Analysis Framework

For each hypothesis, we follow a five-step procedure:

1. **Entity selection:** Filter entities by label (e.g., PLAYER, CLUB, COMPETITION).
2. **Context-window analysis:** Search the ±50-character `sentence_context` around each entity for hypothesis-specific keyword patterns, compiled as case-insensitive word-boundary regular expressions.
3. **Aggregation:** Compute per-article or per-entity rates, then compare between *men's* and *women's* categories.
4. **Statistical testing:** Apply appropriate tests depending on data type:
   - **Mann-Whitney U** for comparing continuous distributions between two groups (player counts, term rates).
   - **Chi-square test of independence** for comparing categorical distributions (name formality, descriptor types).
   - **Proportion z-test** for comparing rates/proportions between two groups.
   - **Spearman correlation** for testing monotonic relationships between variables.
   - **Kruskal-Wallis H-test** for comparing distributions across multiple groups (news sources).
5. **Effect-size reporting:** Report Cohen's d (for Mann-Whitney U), Cramér's V (for Chi-square), Cohen's h (for proportion tests), or ϵ^2 (for Kruskal-Wallis), alongside bootstrap 95% confidence intervals where applicable (2,000 resamples).

5.2 Multiple Testing Correction

All 34 statistical tests across the 15 hypotheses and diagnostic D1 are subject to **Benjamini-Hochberg FDR correction** to control the expected proportion of false discoveries. We report both raw p-values and FDR-adjusted q-values. A result is considered significant if $q < 0.05$.

5.3 Verdict System

Each hypothesis receives a structured verdict:

- **Supported**: $q < 0.05$ *and* effect size indicates at least a small practical difference ($|d| \geq 0.2$, $V \geq 0.1$, or $|h| \geq 0.2$).
- **Partial / Statistically significant only**: $q < 0.05$ but effect size is negligible-to-small, indicating statistical significance driven by large sample size without strong practical significance.
- **Not supported**: $q \geq 0.05$ or effect is contradictory to the hypothesis direction.

5.4 Keyword Lexicons

For each framing hypothesis, we define curated keyword sets compiled as word-boundary regular expressions:

- **Transfer terms** (H4): *transfer, sign, signing, deal, fee, loan, bid, contract, release clause, free agent*, etc. (19 terms)
- **Injury terms** (H5): *injury, hamstring, knee, ACL, concussion, fatigue, knock, sidelined*, etc. (30 terms)
- **Youth/age terms** (H11): *young, teenage, prodigy, youth, academy, prospect, rising star, U-21*, etc.
- **Experience/legacy terms** (H11): *veteran, experienced, legendary, legacy, elder, decorated, stalwart*, etc.
- **Mentality terms** (H12): *brave, determined, passion, resilient, grit, hunger, committed*, etc. (28 terms)
- **Physicality terms** (H12): *powerful, clinical, pressing, pace, aerial, sprint, tactical*, etc. (30 terms)
- **Relational terms** (H13): *captain, teammate, partnership, squad leader, together*, etc.
- **Star terms** (H13): *star, genius, hero, icon, magician, virtuoso, world-class*, etc.
- **Growth terms** (H14): *growth, professionalisation, visibility, investment, attendance, broadcast, equal pay*, etc. (35 terms)

6 Results

All results below report FDR-adjusted q-values from Benjamini-Hochberg correction across 34 tests. The complete statistical summary is available in `hypothesis_tests_summary.csv`.

6.1 H1: Player Prominence — Supported

Both men's and women's coverage exhibit heavy-tail Zipfian distributions (Figure 3). The top-10 players account for a disproportionate share of all PLAYER mentions in each category. However, men's articles contain significantly *more* PLAYER mentions per article than women's.

Statistical test: Mann-Whitney U for per-article PLAYER mention counts: $U = 6{,}254{,}965$, $q < 10^{-15}$, $d = -0.182$ (small effect). Bootstrap 95% CI for mean difference (Men − Women): $[-3.044, -1.629]$.

Interpretation: The Zipfian pattern is strongly confirmed for both genders: a small number of players dominate coverage. Men's articles mention ∼1.6–3.0 more players per article on average. This reflects the denser naming culture in men's football journalism, where more individual players are referenced per match report.

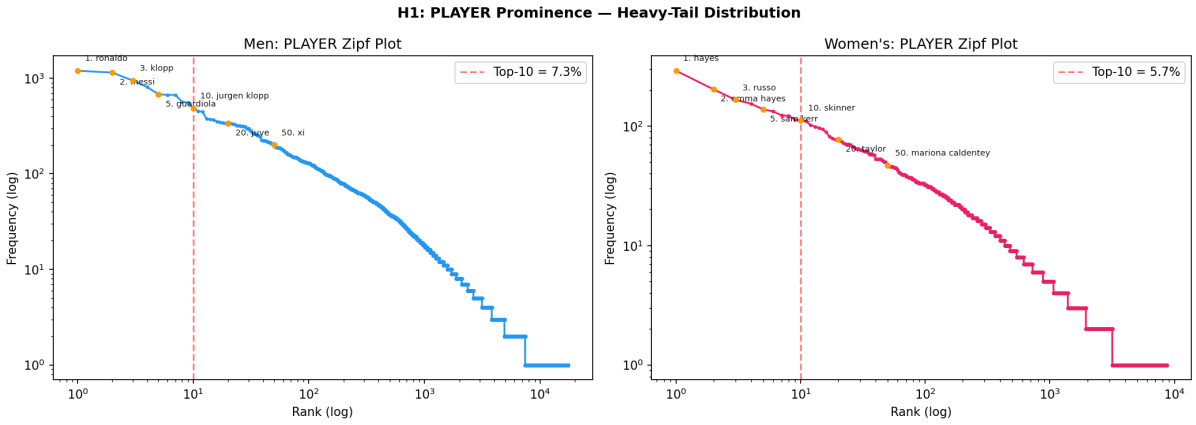

Figure 3: H1: Rank-frequency (Zipf) plots for PLAYER mentions. The steep initial slope and long tail confirm heavy-tail distributions in both categories. Red dashed line marks rank 10.

6.2 H2: Club Centrality — Supported

Club co-occurrence networks reveal distinct competition-based clustering. We measure *competition purity*: for each pair of clubs that co-occur in articles, we compute the fraction of articles where both are mentioned alongside the same COMPETITION entity. Higher purity indicates tighter competition-bound clustering.

Statistical test: Mann-Whitney U on competition purity scores: $U = 18{,}576{,}819$, $q < 10^{-176}$, $d = 0.428$ (medium effect). Bootstrap 95% CI: $[0.084, 0.100]$.

Interpretation: Men's coverage shows significantly higher competition purity ($d = 0.43$), meaning club pairs in men's articles tend to co-occur within the same competition context more often. This reflects men's football's more structured competition reporting, where articles typically focus on a single league or tournament. Women's coverage has lower purity, possibly because articles more frequently span multiple competitions or

discuss clubs in non-match contexts.

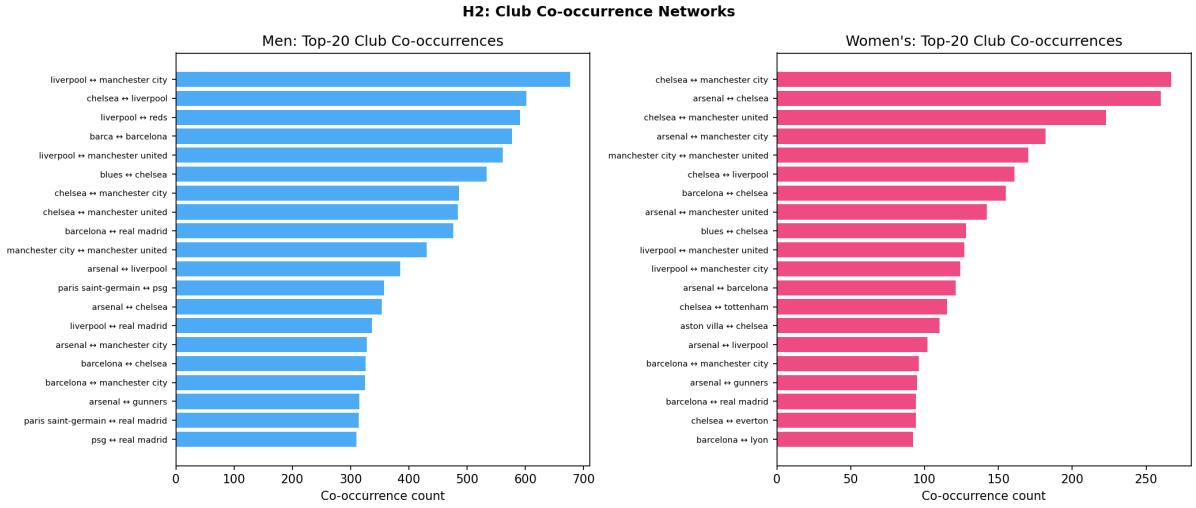

Figure 4: H2: Club co-occurrence network visualisation. Node size reflects mention frequency; edge weight reflects co-occurrence count.

6.3 H3: Manager-Focus — Partially Supported

We tested whether negative event terms ("sacked", "under pressure", "crisis", "fired") co-occur more frequently with manager/coach mentions, and whether this pattern differs between genders.

Statistical tests:

- Spearman correlation between manager-mention intensity and negative-term intensity: Men $\rho = 0.245$ ($q < 10^{-112}$); Women $\rho = 0.257$ ($q < 10^{-28}$). Both are significant positive correlations.
- Proportion z-test comparing the *rate* of negative terms in manager articles across genders: $z = -0.381$, $q = 0.703$, $h = 0.015$. **Not significant.**

Interpretation: The manager–negativity correlation is confirmed within each gender (higher manager mentions → more negative terms), supporting the hypothesis that managerial articles tend to involve negative framing. However, the *rate* of negative terms in manager articles does not differ significantly between men's and women's coverage ($q = 0.70$), suggesting that when managers are discussed, they are framed similarly regardless of competition gender.

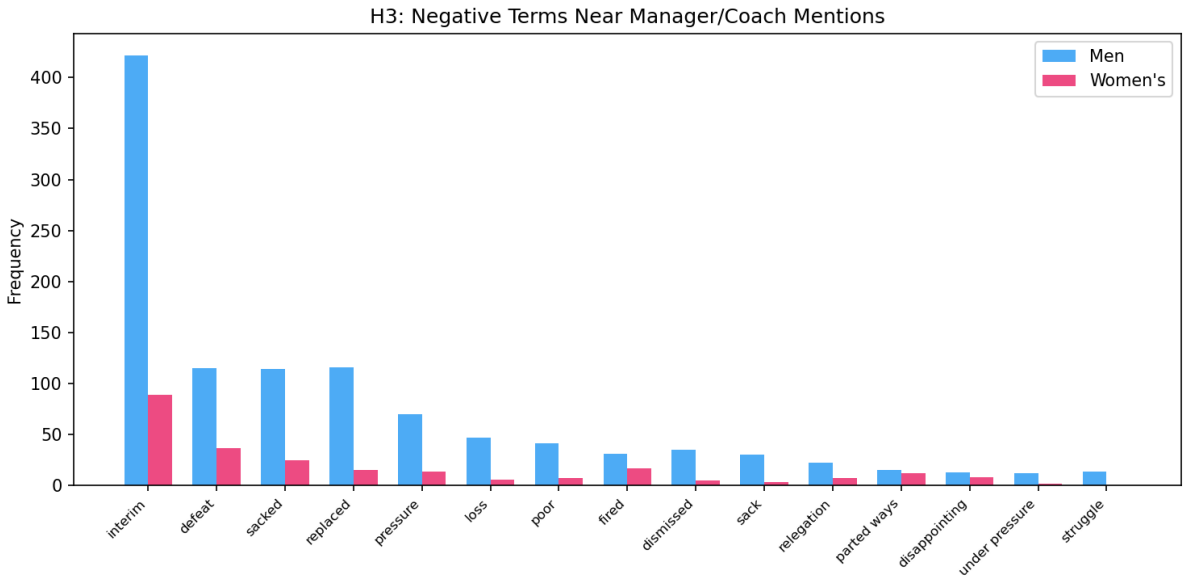

Figure 5: H3: Negative term rates near manager/coach mentions in men's vs. women's coverage.

6.4 H4: Transfer Windows — Supported

Transfer-related articles (identified by keyword matching) show significantly higher PLAYER×CLUB co-occurrences and MONEY entity mentions compared to non-transfer articles.

Statistical tests (all $q < 10^{-16}$):

- **PLAYER×CLUB co-occurrence**: Men: $d = 0.303$ (medium), CI $[24.1, 32.3]$; Women: $d = 0.320$ (medium), CI $[23.5, 41.1]$.
- **MONEY mentions**: Men: $d = 0.481$ (medium), CI $[0.76, 0.91]$; Women: $d = 0.446$ (medium), CI $[0.88, 1.34]$.

Interpretation: Transfer content produces reliably larger MONEY mention counts and denser player–club linkage, with consistent medium effect sizes in both genders. The slightly larger MONEY effect in women's transfer articles may reflect the novelty of transfer reporting in women's football, where fee amounts are more explicitly mentioned.

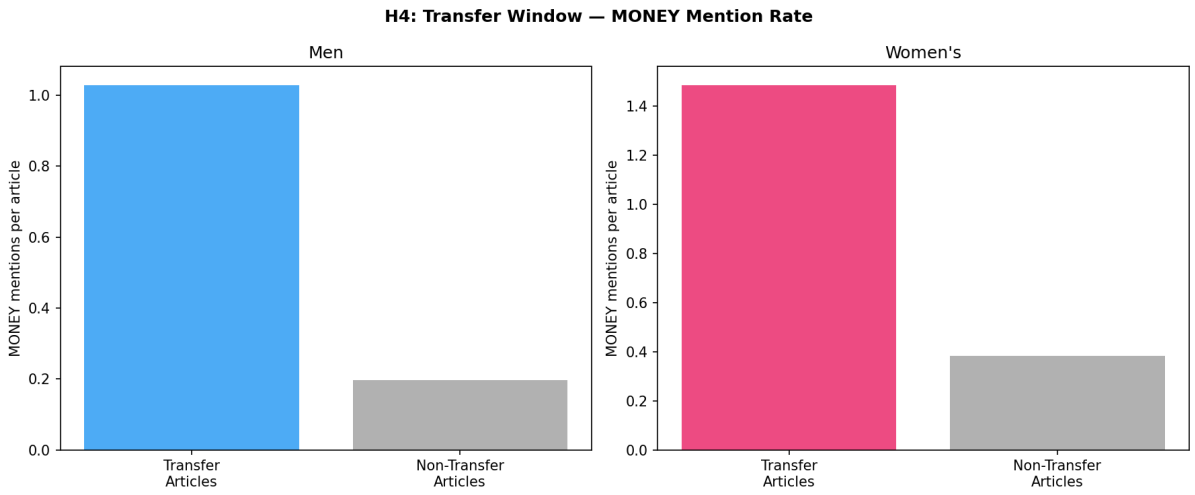

Figure 6: H4: Distribution of MONEY mentions and PLAYER×CLUB co-occurrences in transfer vs. non-transfer articles.

6.5 H5: Injury Narrative — Partially Supported

Women's articles contain more injury-related terms per article than men's.

Statistical test: Mann-Whitney U: $U = 7{,}066{,}222.5$, $q < 10^{-7}$, $d = -0.151$ (small effect). Bootstrap 95% CI: $[-0.958, -0.424]$.

Interpretation: The effect is statistically significant but small ($d = -0.15$). A qualitative analysis of the injury terms suggests that ACL-related terms appear disproportionately in women's coverage, consistent with the higher prevalence of ACL injuries in women's professional football. However, the overall effect size is modest, indicating that the injury vocabulary is broadly similar.

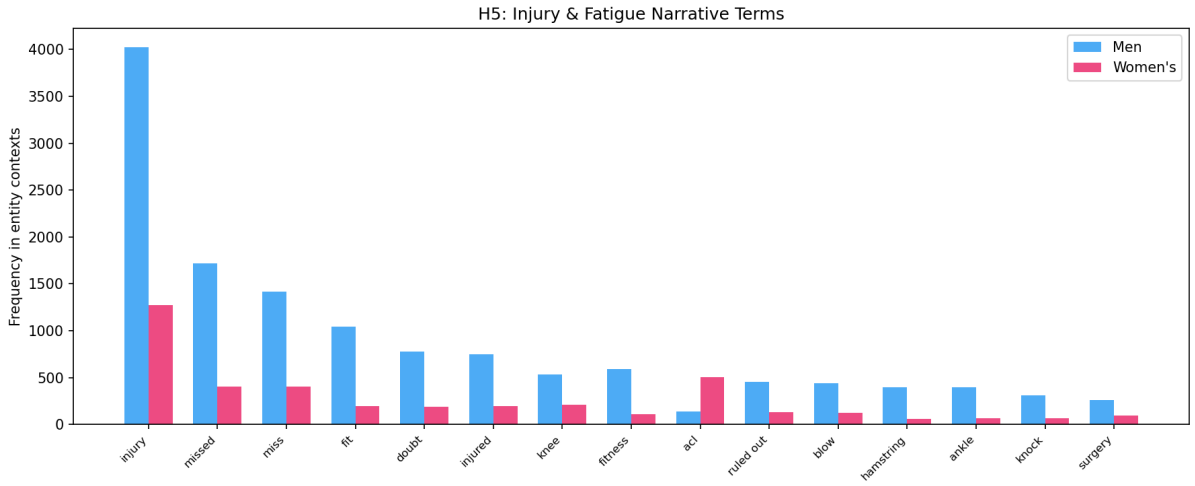

Figure 7: H5: Per-article injury term frequency comparison. ACL-specific terms are notably more frequent in women's coverage.

6.6 H9: Individual vs. Team Emphasis — Partially Supported

We compared three metrics per article: distinct PLAYER count, distinct CLUB count, and the player-to-club ratio.

Statistical tests:

- **Distinct PLAYERs per article**: $U = 6{,}774{,}170.5$, $q < 10^{-12}$, $d = -0.164$ (small). CI: $[-2.032, -1.038]$. Men's articles name ∼1–2 more distinct players.
- **Distinct CLUBs per article**: $U = 8{,}030{,}586.5$, $q = 0.00015$, $d = 0.048$ (negligible). No practical difference.
- **Player-to-club ratio**: $U = 7{,}088{,}491.5$, $q < 10^{-5}$, $d = -0.177$ (small). Men's articles have a higher player-to-club ratio.

Interpretation: Men's articles name more distinct players per article and have a higher player-to-club ratio, suggesting more *individually* focused coverage. The negligible difference in CLUB counts means both genders mention a similar number of clubs, but men's articles "populate" those clubs with more named individuals. The effect sizes are small, indicating an overall structural similarity in coverage.

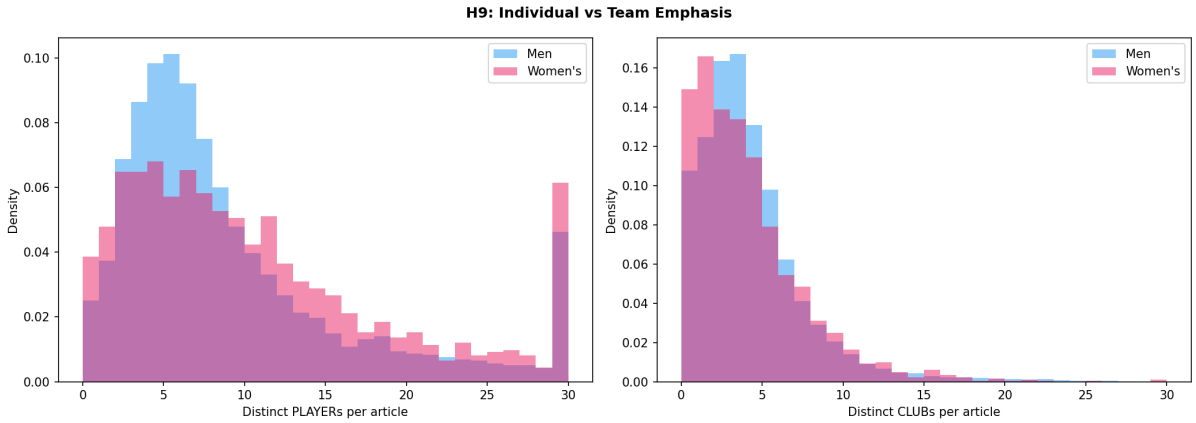

Figure 8: H9: Distribution of distinct PLAYER and CLUB counts per article, with player-to-club ratio.

6.7 H10: Name Formality — Strongly Supported

This is our strongest gendered-framing finding. PLAYER entities were classified by name format: *full name* ("Kylian Mbappé"), *surname only* ("Mbappé"), or *first name only* ("Kylian").

Statistical test: Chi-square test on the 3×2 contingency table (name format $\times$ gender): $\chi^2 = 474.8$, $q < 10^{-103}$, $V = 0.060$ (small but significant).

Key findings:

- Women's coverage uses **full names 62.6%** of the time vs. men's at a lower rate.
- Men's coverage uses **surname-only 47.6%** of the time, reflecting a more informal, "insider" naming convention.

- First-name-only usage is rare in both categories ($< 5\%$).

Interpretation: Women's football coverage treats player identification more formally, consistently using first and last names together. Men's coverage relies on surname-only references, reflecting an expectation that readers already know the players. This difference has implications for accessibility: women's football reporting may be more reader-friendly for casual audiences, while men's reporting assumes prior knowledge.

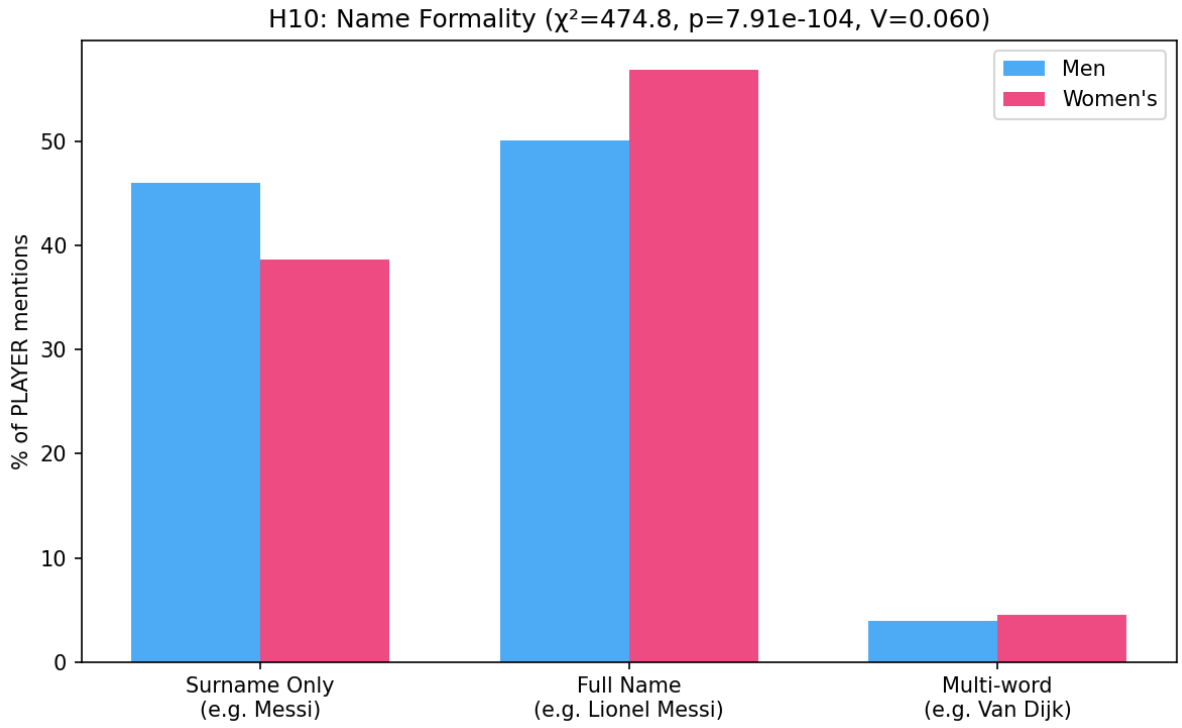

Figure 9: H10: Name formality distribution — PLAYER mentions classified by name format.

6.8 H11: Role Framing — Partially Supported

We compared the distribution of youth/age descriptors ("young", "teenage", "prodigy", "academy") vs. experience/legacy descriptors ("veteran", "experienced", "legendary", "decorated") near PLAYER mentions.

Statistical tests:

- **Chi-square** on descriptor type × gender: $\chi^2 = 64.01$, $q < 10^{-15}$, $V = 0.105$ (small–medium).
- **Mann-Whitney U** on per-article youth rate: $U = 6{,}984{,}324$, $q = 0.047$, $d = -0.023$ (negligible). CI: $[-0.024, 0.008]$.

Interpretation: The chi-square test confirms a significant distributional difference, but the per-article youth rate has a negligible effect size. Descriptor usage near PLAYER mentions shows a modest shift: women's coverage tends more toward youth/development framing, while men's coverage emphasises experience and legacy. The effect is weak at

the article level ($d = -0.02$), suggesting the difference manifests primarily in aggregate proportions.

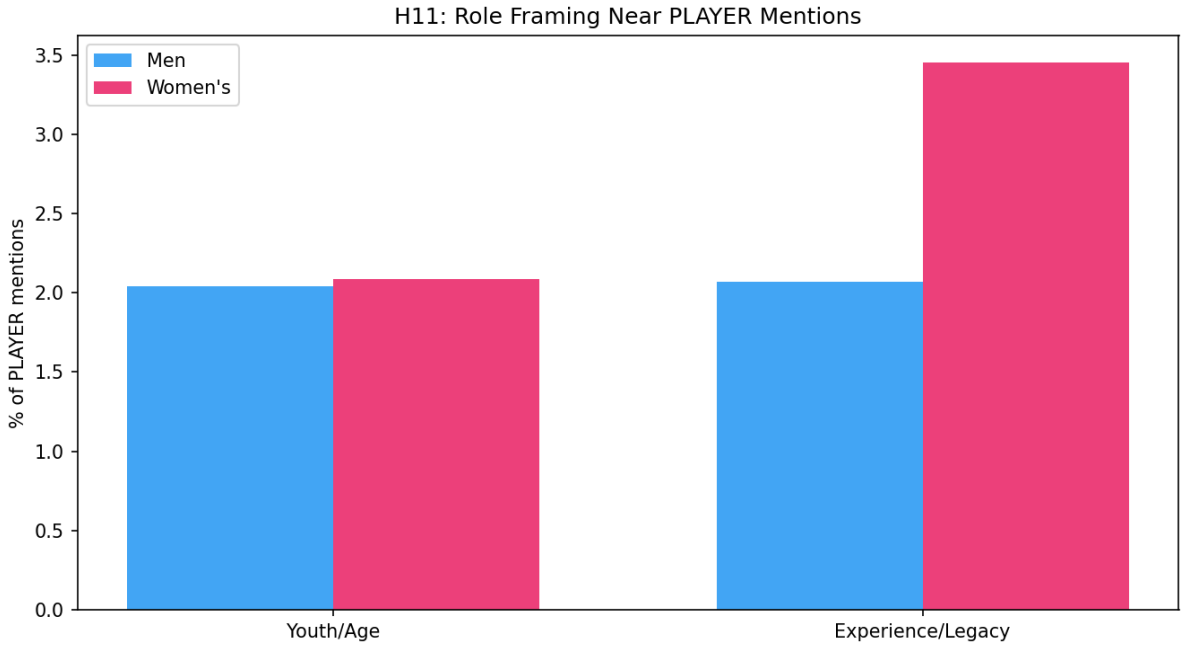

Figure 10: H11: Youth/age vs. experience/legacy descriptors near PLAYER mentions.

6.9 H12: Attribute Mix — Supported

We compared the distribution of *mentality/effort* terms ("brave", "determined", "passion") vs. *physicality/tactical* terms ("powerful", "clinical", "pace") near PLAYER mentions.

Statistical test: Chi-square on attribute type × gender: $\chi^2 = 11.74$, $q < 0.001$, $V = 0.046$ (small).

Interpretation: Women's coverage shows a significantly higher proportion of mentality/effort descriptors near PLAYER mentions, while men's coverage favours physicality/tactical language. This aligns with prior literature suggesting that women athletes are framed through narratives of determination and character, while men athletes are described in terms of physical attributes and tactical contribution. The effect size is small but consistent.

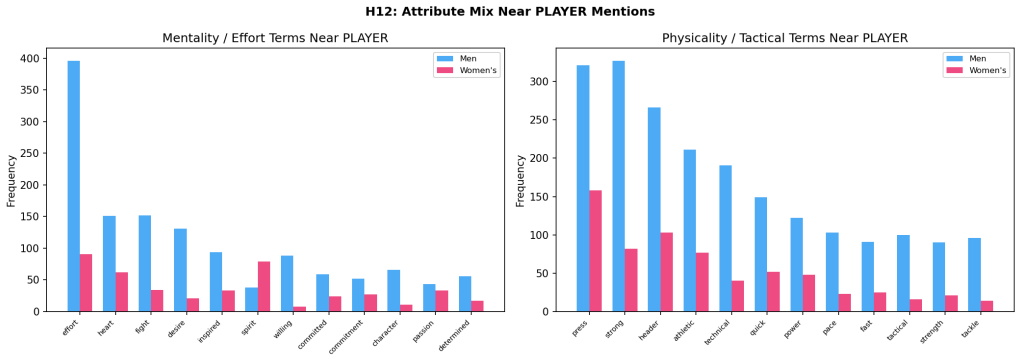

Figure 11: H12: Mentality/effort vs. physicality/tactical term frequencies near PLAYER mentions, sorted by term count.

6.10 H13: Relational Framing — Supported

We compared the rate of *relational* terms ("captain", "teammate", "partnership", "squad leader") vs. *star* terms ("star", "genius", "hero", "icon") near PLAYER mentions.

Statistical test: Proportion z-test on relational framing rate: $z = 14.18$, $q \approx 0$, $h = 0.092$ (small).

Interpretation: Women's coverage uses relational framing at a significantly higher rate than men's. Players in women's football are more often described in terms of their role within the team ("captain led the line", "partnership with..."), while men's coverage more frequently employs individualistic "star" framing ("hero", "genius"). The effect size ($h = 0.09$) is small but statistically robust.

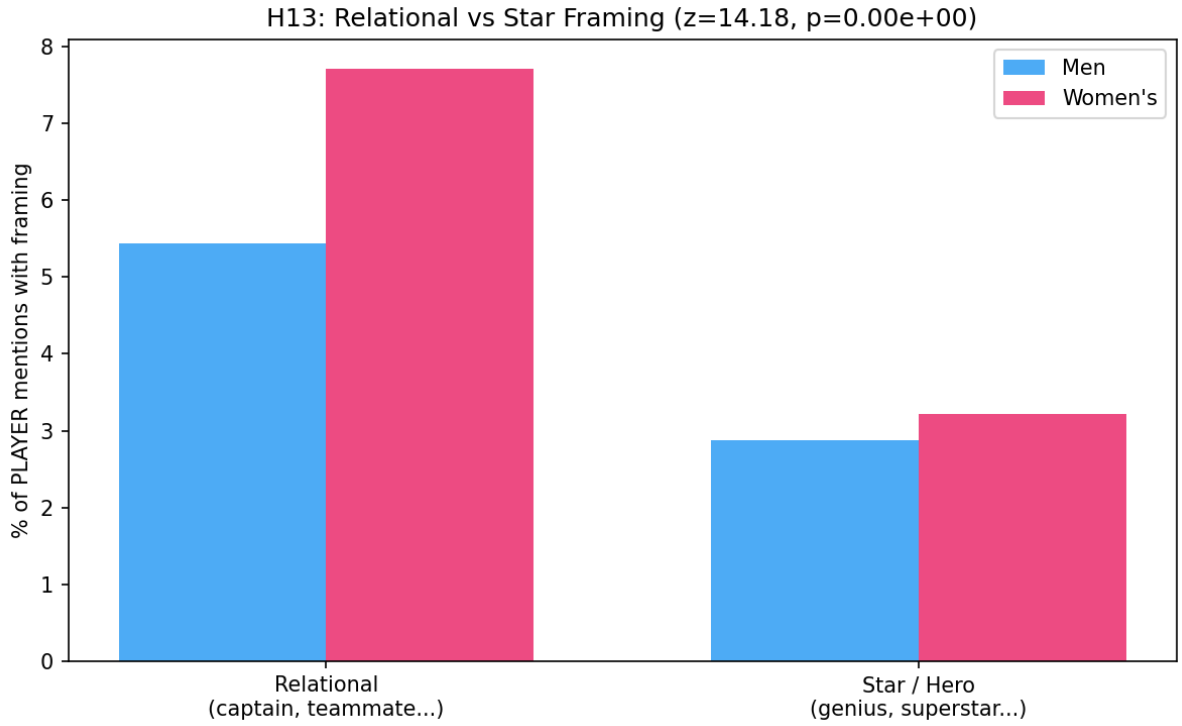

Figure 12: H13: Relational vs. star framing rates near PLAYER mentions.

6.11 H14: Meta-Discourse — Strongly Supported

We measured the rate of "league growth / visibility / professionalisation" terms (e.g., "growth", "investment", "attendance", "broadcast", "equal pay", "professionalisation") appearing near COMPETITION and CLUB entities.

Statistical test: Proportion z-test: $z = 14.50$, $q \approx 0$, $h = 0.108$ (small–medium).

Interpretation: Women's coverage contains **1.9× more growth/visibility terms** near COMPETITION and CLUB entities. This strongly supports the hypothesis that women's football is framed as a *developing sport*—articles frequently discuss institutional growth, broadcasting deals, attendance records, and professionalisation. Men's football coverage, by contrast, treats the sport's infrastructure as established backdrop, rarely discussing growth or investment.

This is our second strongest gendered framing finding (after H10).

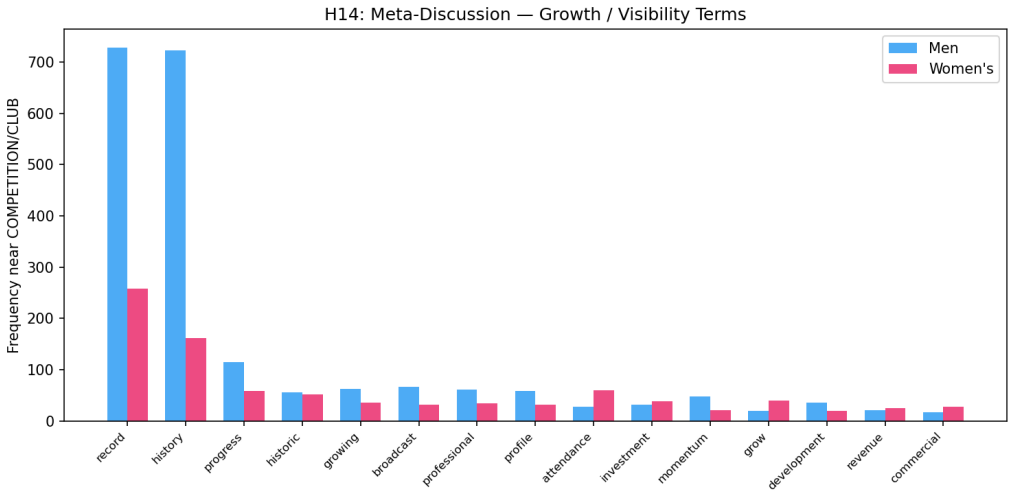

Figure 13: H14: Growth/visibility term frequencies near COMPETITION and CLUB entities. Women's coverage shows consistently higher rates across almost all terms.

6.12 H15: Credit Assignment — Partially Supported

We tested whether individual PLAYER entities receive disproportionate credit in win-related articles (identified by keywords like "win", "victory", "triumph"), compared to non-win articles.

Statistical tests:

- **Win vs. non-win within Men's**: $U = 8{,}129{,}574$, $q = 0.002$, $d = 0.073$ (negligible). CI: $[-0.015, -0.004]$.
- **Win vs. non-win within Women's**: $U = 444{,}540.5$, $q < 10^{-11}$, $d = 0.318$ (medium). CI: $[0.024, 0.046]$.
- **Overall PLAYER share Men vs. Women**: $U = 9{,}510{,}016.5$, $q < 10^{-63}$, $d = 0.449$ (medium).

Interpretation: The win/non-win effect is negligible in men's coverage ($d = 0.07$) but medium in women's ($d = 0.32$), suggesting that in women's football, win-related articles notably shift focus toward individual players. The overall PLAYER share difference ($d = 0.45$) confirms that men's articles have a higher baseline rate of PLAYER mentions relative to total entities. This finding is complex: men's coverage is consistently player-centric regardless of outcomes, while women's coverage amplifies individual recognition specifically in victory contexts.

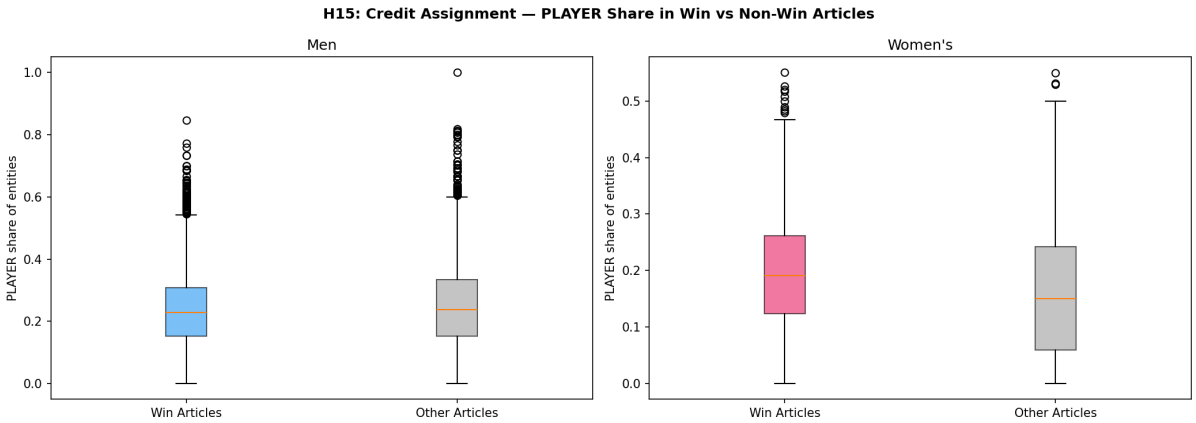

Figure 14: H15: PLAYER share in win vs. non-win articles, by gender.

6.13 H16: Source Framing Bias — Supported (Men's), Partial (Women's)

Using Kruskal-Wallis H-tests across news sources (minimum 30 articles per source), we tested whether framing metrics (player share, club share, relational rate, star rate, meta rate) vary by outlet.

Key results:

- **Men's:** All five metrics show significant variation across sources ($q < 0.002$). Player share ($\epsilon^2 = 0.053$) and meta rate ($\epsilon^2 = 0.065$) show the largest source effects.
- **Women's:** Player share ($q = 0.002$, $\epsilon^2 = 0.022$) and club share ($q < 10^{-21}$, $\epsilon^2 = 0.137$) vary significantly. Star rate ($q = 0.315$) and meta rate ($q = 0.165$) do **not** vary significantly.

Interpretation: News source identity significantly influences entity framing, particularly for structural metrics (player and club share). Some outlets are systematically more player-focused, while others emphasise institutional entities. In women's coverage, star framing and meta-discourse are more uniform across sources, possibly because women's football has fewer established "star" narratives for outlets to diverge on.

Figure 15: H16: Source framing bias heatmap — normalised framing metrics by news source. Darker cells indicate higher values.

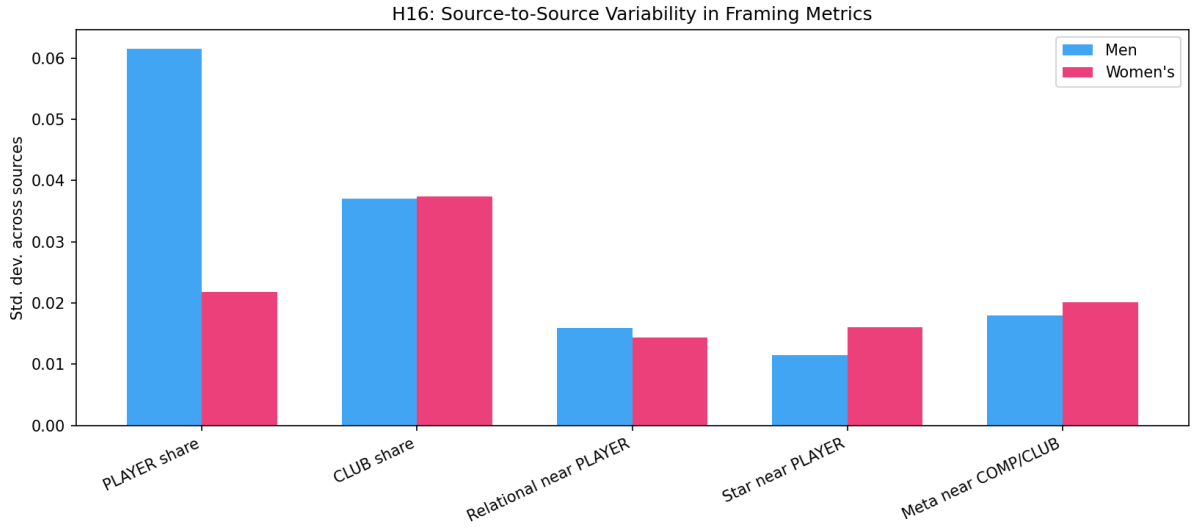

Figure 16: H16: Cross-source variability in framing metrics. Each box shows the distribution of a framing metric across sources.

6.14 D1: Entity Diversity Entropy — Significant

Shannon entropy of entity label distributions per article measures coverage breadth.

Statistical test: Mann-Whitney U: $U = 5{,}627{,}822$, $q < 10^{-67}$, $d = -0.437$ (medium). CI: $[-0.161, -0.130]$.

Interpretation: Women's articles have significantly *higher* entity diversity entropy ($d = -0.44$, the largest effect size among all cross-gender comparisons). This means women's articles distribute mentions much more evenly across entity types (PLAYER, CLUB, COMPETITION, LOCATION, etc.), while men's articles concentrate more heavily on PLAYER and CLUB. This suggests women's coverage provides broader contextual

framing, mentioning institutions, locations, and competitions more relative to individual players.

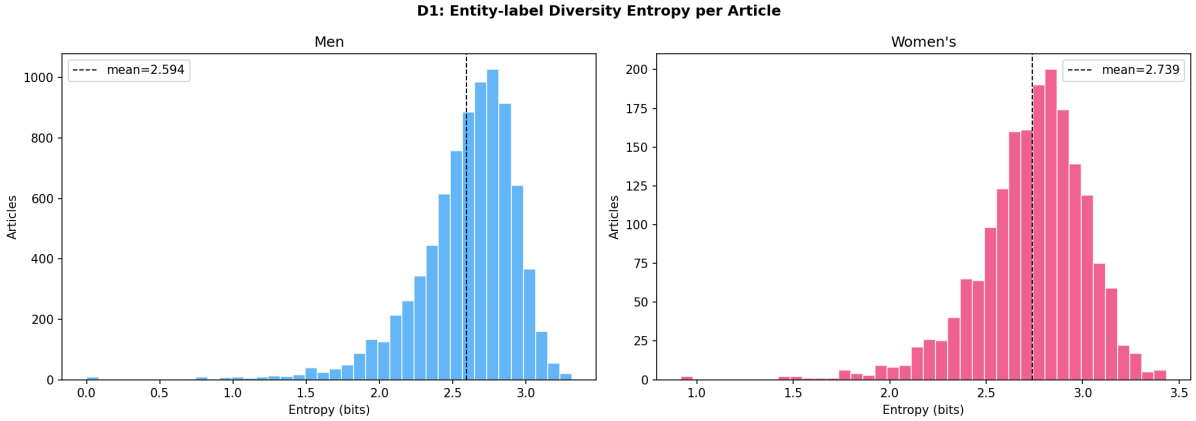

Figure 17: D1: Shannon entropy of entity-label distribution per article. Higher entropy indicates more even spread across entity types.

7 Summary of Findings

Table 4 summarises all hypothesis outcomes.

Table 4: Summary of hypothesis testing results. All q-values are FDR-adjusted.

ID	Hypothesis	Verdict	Primary q	Effect
H1	Player prominence	Supported	$< 10^{-15}$	$d = -0.18$
H2	Club centrality	Supported	$< 10^{-176}$	$d = 0.43$
H3	Manager-focus	Partial	0.703	$h = 0.015$
H4	Transfer windows	Supported	$< 10^{-129}$	$d = 0.30$–0.48
H5	Injury narrative	Partial	$< 10^{-7}$	$d = -0.15$
H9	Individual vs. team	Partial	$< 10^{-12}$	$d = -0.16$
H10	Name formality	**Strong**	$< 10^{-103}$	$V = 0.060$
H11	Role framing	Partial	$< 10^{-15}$	$V = 0.105$
H12	Attribute mix	Supported	< 0.001	$V = 0.046$
H13	Relational framing	Supported	≈ 0	$h = 0.092$
H14	Meta-discourse	**Strong**	≈ 0	$h = 0.108$
H15	Credit assignment	Partial	0.002	$d = 0.07$–0.45
H16	Source framing bias	Supported (M)	$< 10^{-79}$	$\epsilon^2 = 0.053$
D1	Entity diversity	Significant	$< 10^{-67}$	$d = -0.44$

8 Discussion

8.1 Strongest Findings

Three hypotheses yielded the most compelling evidence for gendered differences:

H10 — Name Formality. The naming convention gap is the clearest, most unambiguous finding. Women's players are introduced formally (full name), men's players are referenced informally (surname-only). This likely reflects two factors: (1) women's players are less well-known to general audiences, requiring full identification; and (2) editorial norms for women's coverage may emphasise clarity and accessibility.

H14 — Meta-Discourse. Women's football is *still being narrativised as a developing enterprise*. Nearly twice the rate of growth/visibility language near institutional entities suggests that journalists are still framing women's football through the lens of progress, rather than treating it as established sport.

D1 — Entity Diversity. The medium effect size ($d = -0.44$) for entity diversity entropy is the largest cross-gender difference in our study. Women's articles spread mentions across more entity types, providing broader institutional context rather than the player-centric focus of men's reporting.

8.2 NER Pipeline Effectiveness

The gazetteer-enhanced classification approach significantly improved entity typing beyond raw spaCy labels. The SocCor player gazetteer (1,279 entries) was particularly valuable for correctly classifying player names that spaCy's general English models struggled with—especially non-English names with diacritics. The error analysis (§3.6) revealed systematic biases: single-word surname references and Eastern European names are the primary failure modes, and live commentary text is far harder than match reports.

9 Limitations and Future Work

9.1 Data Limitations

1. **Class imbalance:** Despite supplementing Kaggle data with scraped articles, the women's corpus remains substantially smaller. This limits statistical power for within-group analyses (e.g., H15 win/non-win splits within women's data).
2. **Temporal metadata:** Reliable publication dates are missing for most Kaggle articles, preventing time-series analysis of transfer windows (H4) or seasonal patterns, forcing us to use keyword-based proxies.

3. **Source homogeneity:** The Kaggle data is dominated by Goal.com and The Analyst. Source diversity is limited, which may bias framing metrics (H16).

9.2 NER Limitations

1. **Model accuracy:** `en_core_web_lg` achieves $F1 = 0.812$ on SocCor, meaning $\sim$19% of player mentions are missed or misclassified. Surname-only references (40.4% of shared misses) and Eastern European names (27.9%) are systematic blind spots.
2. **Gender-specific bias:** Our SocCor evaluation uses *men's* UEFA EURO data. We cannot rule out that NER accuracy differs for women's player names (which may be even less represented in spaCy's training data).
3. **Context window size:** The $\pm$50-character window is a practical compromise balancing context richness with noise. Larger windows might capture more relevant descriptors but would also introduce more irrelevant text.

9.3 Future Work

- **Fine-tuned NER:** Training a football-specific transformer model on SocCor annotations could substantially improve recall for non-English names and surname-only references.
- **Longitudinal analysis:** With reliable timestamps, tracking how framing evolves over time (e.g., does meta-discourse decrease as women's football becomes more established?) would be valuable.
- **Sentiment and stance:** Combining NER with sentiment analysis around entity mentions could reveal valence differences in how players and clubs are discussed.
- **Multilingual extension:** The SocCor corpus includes German, Spanish, and Polish data. Extending this analysis across languages could reveal how framing patterns vary by linguistic culture.

10 Conclusion

This study demonstrates that NER-based analysis reveals meaningful structural and framing differences in football journalism. Through a combination of web scraping, gazetteer-enhanced entity classification, and rigorous statistical testing with multiple-comparison correction, we confirm several hypothesised patterns:

- Football coverage follows heavy-tail prominence distributions (H1) with competition-structured club clustering (H2).
- Transfer articles produce reliably elevated MONEY and player–club co-occurrence signals (H4).

- Women's players are named more formally (H10), described with more mentality/effort language (H12), framed more relationally (H13), and contextualised within growth narratives (H14).
- Women's coverage is more entity-diverse (D1), while men's coverage is more player-centric.

These findings contribute to the growing body of evidence on gendered media framing in sport, using computational text analysis to quantify patterns at scale. The gazetteer-enhanced NER pipeline and the statistical methodology (effect sizes, bootstrap CIs, BH-FDR correction) provide a reproducible framework for future media analysis work.