④ 名詞由来エッジ抽出プロンプト（幕/場スコープ対応・感情なし・代名詞除外・エラー行出力あり）

あなたは、指定スコープ（幕/場）の脚本テキストから、

「名詞（固有名／役称／綴りゆれ等）に基づく有向辺」を抽出し、CSVのみを出力します。

※ 代名詞は対象外（代名詞は別プロンプトで処理）。

# 出力ヘッダ（順序固定・厳密一致）

Edge\_Id,Work\_Name,Speech\_Id,Source\_Node\_Id,Target\_Node\_Id,Alias\_Id,Alias\_Label,Span\_Start,Span\_End,Matched\_By,Excluded\_Flag,Speech\_Quote,Remarks

# 入力

<WORK\_NAME>

（作品タイトル）

</WORK\_NAME>

<SCOPE>

Unit: act | scene

Value: 例) I または I-3

</SCOPE>

<NODE\_LIST\_CSV>

Node\_Id,Name,Work Name,First Line,Grammatical Number,Existed on initial list?,Remarks

…（このスコープを含むノードリストCSV）…

</NODE\_LIST\_CSV>

<ALIAS\_LIST\_CSV>

Work\_Name,Alias\_Id,Node\_Id,Node\_Name,Alias\_Label,Alias\_Type,Case\_Sensitive,Is\_Regex,Whole\_Word,POS,Disambiguation\_Rule,Unique\_in\_Scope,Candidate\_Node\_Ids,Used\_in\_Speech,First\_Attested\_Speech\_Id,Evidence\_Quote,Exclude\_from\_Edges,Remarks

…（呼称リストの出力；※代名詞は含まれない）…

</ALIAS\_LIST\_CSV>

<SCOPE\_SCRIPT>

（指定スコープ＝該当の幕/場のみの原文。発話ラベルと本文を含む）

</SCOPE\_SCRIPT>

# 制約

- CSVのみ出力（解説・空行・コードフェンス禁止／RFC 4180でエスケープ）。

- 代名詞は本プロンプトの対象外。

- Wikisource等のブロック番号は使用しない。位置は Speech\_Id + Span（発話内オフセット）で特定。

- 1一致 = 1行（同一発話で同一対象に複数出現すれば行を複数出力）。

# 手順（要点）

1) 発話抽出と Speech\_Id 付与

- <SCOPE> に従い当該スコープのみ処理。

- <SCOPE\_SCRIPT> から「話者ラベル＋本文」を列挙し、Scene単位で Speech\_Id を採番。

形式：Act-Scene-U連番（例：I-3-U0007）。Sceneが無い場合は Act 内連番（例：I-U0007）。

2) 話者（Source\_Node\_Id）の同定

- 話者ラベルを <NODE\_LIST\_CSV> に突合して Source\_Node\_Id を確定。

- グループ話者は個体に一意分解できる場合のみ各個体に複製（分解不能なら通常出力は生成しないが、後述「6-b エラー行」を適用）。

3) 名詞呼称候補の確定

- <ALIAS\_LIST\_CSV> から下記を満たす行のみ「通常出力の候補集合」とする：

- Exclude\_from\_Edges = n

- Unique\_in\_Scope = y

- これらはすべて名詞系エイリアス（代名詞は含まれない）。

4) マッチング

- 各発話本文に対し、候補行の設定を厳格に適用して Alias\_Label 照合：

- Case\_Sensitive（既定 n）

- Is\_Regex（既定 n。y の場合は安全条件：Unicodeモード必須／貪欲量指定 `.\*` 不可／量指定は最大40文字程度／語境界は `\b` のみ頼りにせず空白・句読点・行端も許容）

- Whole\_Word（既定 y。Regex時はパターン側で表現）

- Disambiguation\_Rule（`speaker\_in|near\_any|window|scene|not\_near|note` の最小条件。過解釈禁止）

- 一致した箇所ごとに：

- Target\_Node\_Id = 該当 Alias 行の Node\_Id

- Matched\_By = alias-exact / alias-regex / alias-normalized

- Span\_Start / Span\_End = 発話本文内の0始まり文字オフセット（Endは排他）

- Speech\_Quote = 該当箇所を含む最大10語の抜粋（CSV安全）

- Excluded\_Flag = n（通常行）

5) 重複・衝突処理

- 同一 `Speech\_Id / Span\_Start / Span\_End / Target\_Node\_Id / Alias\_Id` の完全重複は1行に統合。

- 同一スパンに複数の Alias が衝突する場合は通常行を生成せず、6-b のエラー行を生成（Remarks="エイリアス衝突"）。

6) 除外・エラー行の生成（Excluded\_Flag = y）

- 下記に該当する場合は\*\*エラー行\*\*を生成し、Excluded\_Flag=y、Remarks に理由を記す。

a) 話者ノード特定不能：`Source\_Node\_Id` を空欄、`Remarks="話者特定不能"`

b) 集合話者分解不能：`Source\_Node\_Id` に集合ノードID、`Remarks="集合話者・個体分解不能"`

c) 曖昧呼称でターゲット一意化失敗（Unique\_in\_Scope=n 等）：

- `Target\_Node\_Id` 空欄

- `Remarks="曖昧（候補: Nxx;Nyy）"` のように候補IDを `;` 区切りで列挙

- `Matched\_By` は実際の一致種別（alias-exact / alias-regex / alias-normalized）

d) 除外規則適用（例：Used\_in\_Speech=n、呼称タイプが group\_label 等）：

- `Remarks="除外規則適用(Used\_in\_Speech=n)"` など具体的に記載

e) エイリアス衝突（同一スパンで複数一致）：

- `Remarks="エイリアス衝突"`

- エラー行でも `Alias\_Id / Alias\_Label / Span\_\* / Speech\_Quote / Matched\_By` は記録する（特定できた範囲で）。

7) Edge\_Id（主キー）の生成（決定的・通常/エラー共通）

- 通常行の形式（推奨）：

{WorkSlug}|{Speech\_Id}|S{Source\_Node\_Id}|T{Target\_Node\_Id}|{Alias\_Id}|X{Span\_Start}-{Span\_End}

- エラー行のプレースホルダ規約：

- 話者不明：`SNA` を使用（例：`SNA`）

- ターゲット不明：`TNA` を使用（例：`TNA`）

→ 例：{WorkSlug}|{Speech\_Id}|SSNA|TTNA|{Alias\_Id}|X{Span\_Start}-{Span\_End}

- WorkSlug：Work\_Name を小文字ASCIIに正規化（ä→ae, ö→oe, ü→ue, ß→ss、記号除去、先頭12文字程度）

8) 出力

- 次のヘッダで CSV のみ出力（順序・綴り厳守、フィールドにカンマ/改行/引用符がある場合はRFC 4180準拠でエスケープすること）：

Edge\_Id,Work\_Name,Speech\_Id,Source\_Node\_Id,Target\_Node\_Id,Alias\_Id,Alias\_Label,Span\_Start,Span\_End,Matched\_By,Excluded\_Flag,Speech\_Quote,Remarks