日本語文埋め込み空間に現れる空間認知方法表現の 1 次元軸

近藤泰弘
青山学院大学
yhkondo@cl.aoyama.ac.jp

概要

本研究では，空間記憶の方法である egocentric（自己中心）と allocentric（場所中心）の表現の，文埋め込み空間での状態を記述する．空間参照枠を示す語のうち，主に egocentric とされる「前後左右」を含む文と allocentric とされる「東西南北」とを，ベクトル化した上で，主成分分析（PCA）等により分離構造を評価し，PC1 に沿って一次元的に分離することを示す．また，指示詞などの deictic 表現が egocentric 側でのみ顕在化するものとして現れることを示し，空間認知の言語体系が，文埋め込み空間で線形に符号化されている可能性を示す．

1 はじめに

認知言語学において，空間参照枠（frames of reference）における relative frame と absolute frame の対立に典型的である，egocentric（自己中心）と allocentric（場所中心）の区別は，空間認知を規定する基本概念である．一方，近年の transformer に基づく文埋め込みで，こうした参照枠対立がベクトル空間内でどのような幾何学的構造として保持されているかは十分に明らかではない．

本研究は，日本語の空間参照枠語である方向語の体系（東西南北／前後左右）を「参照枠対立の診断子」として用い，文埋め込み空間において egocentric–allocentric の空間認知が 1 次元の線形軸として抽出可能かを検証する．

本研究の手法と目的は次の 3 点である：

- 方向語（東西南北 vs 前後左右）を含む文が，文埋め込み空間で一次元的（PC1）に分離し得ることを示す．
- 背景主成分の診断・除去を含む統制により，その分離が単純な文体差や背景寄生のみでは説明しにくいことを示す．
- direction-PC1 へのアンカー射影により，deixis が egocentric 側で顕在化するものである可能性を示し，ベクトル空間内での，空間認知の立体的な幾何学的構造を明らかにする．

2 関連研究

空間参照枠（frames of reference）の研究 空間参照枠は，相対参照（relative/egocentric）と絶対参照（absolute/allocentric）などの区別として整理され，空間記述・談話解釈に関わる認知言語学／認知科学の基本概念として, 諸言語で研究されてきた（参考文献 [1, 2, 3, 4] 等）

文埋め込みと内部表現の線形可読性 Sentence-BERT 系の文埋め込みや対照学習型テキスト埋め込みは多様な下流タスクで広く用いられ（参考文献 [5] 等），近年はモデル内部表現から特定概念を線形に読み出す（linear probe）研究も進んでいる（参考文献 [6] など）．

3 手法

3.1 文埋め込み

`intfloat/multilingual-e5-large-instruct` を文埋め込みモデルとして用いて，各文から，埋め込みベクトルを得る．なお，以下，特に断りがない限り，埋め込みは L2 正規化を施したものを用いる．

3.2 合成文生成

方向語を含む文は，複数のテンプレートと語彙スロットを用いて自動生成する．本研究では，語彙対立として，allocentric（絶対参照側）の方向語群（例：東西南北）と，egocentric（相対参照側）の方向語群（例：前後左右）を用いる．具体的には「地図の向きで言うと，駐車場の東方に入口がある.」「体感としては，学校は前方にある.」などの形となる．また，テンプレートは文長・句読点・語彙スロット位置を一定範囲で制御できるよう設計した．

4 手法

4.1 PCA と分離指標

埋め込み行列 $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times d}$ に対し，平均中心化（mean centering）を施した後に主成分分析（PCA）を適用し，主成分ベクトル（固有ベクトル）$\{u_k\}_{k \geq 1}$ と各文の主成分スコアを得る．特に第 1 主成分（PC1）スコア $z_{i,1}$ とラベル y_i（例：egocentric/allocentric）の関係に注目する．

効果量（η^2）　PC1 スコアの分散のうち，ラベル群間変動で説明される割合として ANOVA の η^2 を用いる．

プローブベクトルによる解釈（Direction Probe） PC1 軸の意味解釈を行うため，方向語群の平均埋め込み差としてプローブベクトルを定義し，PC1 方向との整列度（alignment）を評価する．Egocentric 語彙集合を $\mathscr{E}$，Allocentric 語彙集合を $\mathscr{A}$ とし，各集合の平均埋め込みを以下で定義する．

$$\mu_E = \frac{1}{|\mathscr{E}|} \sum_{w \in \mathscr{E}} \bar{x}(w), \quad \mu_A = \frac{1}{|\mathscr{A}|} \sum_{w \in \mathscr{A}} \bar{x}(w) \tag{1}$$

ここで $\bar{x}(w)$ は，語 w を含む文埋め込みの平均ベクトルとする．このとき，Egocentric と Allocentric の差分ベクトルを

$$v_{\text{probe}} = \mu_E - \mu_A \tag{2}$$

と定義し，これとデータ PC1 単位ベクトル u_1 とのコサイン類似度の絶対値を整列度とする．

$$\text{align}(v_{\text{probe}}, u_1) = |\cos(v_{\text{probe}}, u_1)| = \frac{|v_{\text{probe}}^\top u_1|}{\|v_{\text{probe}}\|_2 \|u_1\|_2} \tag{3}$$

絶対値を用いるのは，PCA の軸方向の符号が任意であるためである．

4.2 2×2 要因計画（Frame × Token）

方向語の分離が文脈フレームに依存するのか，語彙対立そのものが支配的かを切り分けるため，文全体としてのフレーム $a \in \{\text{ego}, \text{allo}\}$ と方向語の語類 $b \in \{\text{rel}, \text{abs}\}$ の 2×2 要因で合成文を生成する（rel=前後左右，abs=東西南北）．各セルの平均埋め込みベクトルを $\mu_{a,b}$ とし，フレームごとの語彙対立ベクトル（Ego 方向への差分）を以下のように定義する．

$$d_{\text{ego}} = \mu_{\text{ego,rel}} - \mu_{\text{ego,abs}}, \quad d_{\text{allo}} = \mu_{\text{allo,rel}} - \mu_{\text{allo,abs}} \tag{4}$$

これらを用いて，語彙の主効果（Global Axis）およびフレームとの交互作用（Interaction）を以下のベクトルとして算出する．

$$\Delta_{\text{token}} = \frac{d_{\text{ego}} + d_{\text{allo}}}{2}, \quad v_{\text{int}} = \frac{d_{\text{ego}} - d_{\text{allo}}}{2} \tag{5}$$

4.3 背景 PC 診断・除去（Background-PC Removal）

「方向語の軸が，言語データ一般に含まれる支配的な主成分（文長や頻度など）に寄生しているだけではないか」という疑念（疑似相関）を検証する．背景文集合 $\mathscr{B} = \{t_j\}_{j=1}^M$ を同一モデルで埋め込み $b_j \in \mathbb{R}^d$ に変換し，これに PCA を適用して上位 K 本の背景主成分 $\{u_k^{(\text{bg})}\}_{k=1}^K$ を得る．データ側の注目ベクトル v（例：データ PC1 や Δ_{token}）と背景 PC の整列度は以下で評価する．

$$a_k = |\cos(v, u_k^{(\text{bg})})| \tag{6}$$

さらに，各データ埋め込み x_i から背景 PC 成分を除去（Rejection）し，補正済みベクトル x_i' を得る．

$$x_i' = x_i - \sum_{k=1}^{K} (x_i^\top u_k^{(\text{bg})}) u_k^{(\text{bg})} \tag{7}$$

本研究では K の値（除去本数）を変化させ（例：rm1, rm5, rm10），結果の頑健性を確認する．

4.4 方向語 PC1 へのアンカー射影

方向語が作る global な ego/allo 二極軸と，他のダイクティック表現が示す局所差分を同一の座標系で比較するため，方向語データのみから推定した PC1 を **アンカー軸**として固定し，他データセットをその軸へ射影する．

具体的には，方向語データ集合 $\mathscr{D}_{\text{dir}}$ の平均 μ_{dir} と，その PC1 単位ベクトル u_{dir} を得る．任意の文埋め込み x に対し，direction-PC1 への射影スコアを

$$s_{\text{dir}}(x) = (x - \mu_{\text{dir}})^\top u_{\text{dir}} \tag{8}$$

と定義する．符号は解釈のために固定し，相対参照（前後左右；egocentric）側の平均スコアが正になるよう u_{dir} の向きを反転する．

この s_{dir} により，方向語（東西南北 vs 前後左右）が 0 をまたいで分布する（bipolar/global）のに対し，「〜ていく／〜てくる」「授受（〜てやる／〜てくれる／〜てもらう）」「こそ（こ／そ）」等が主に egocentric 側半空間へ偏る（polarize/within-ego）という差を可視化・定量化する．

5 実験と結果

本節では，(i) 合成文で観察される方向語の線形分離，(ii) 交絡要因・背景主成分を考慮しても崩れに

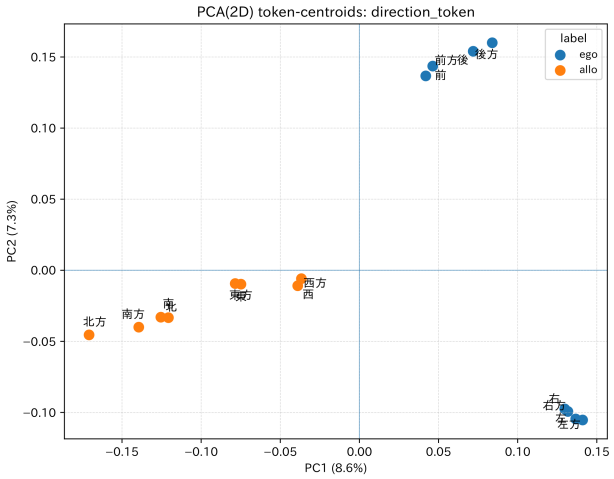

図 1　方向語を含む合成文ベクトルの PCA2 次元散布図・左が東西南北，右が前後左右

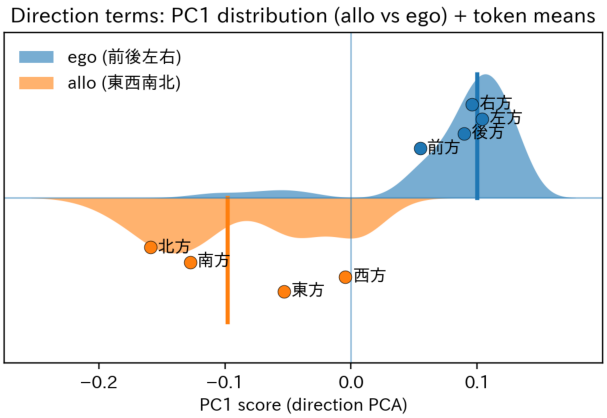

図 2　方向語を含む合成文ベクトルの PCA の PC1 での分布図

くい頑健性，(iii) 方向語で定義した軸に対するダイクティック表現の偏在性，を順に示す．

5.1　方向語：合成文・自然文における PC1 分離

方向語（allocentric: 東西南北／ egocentric: 前後左右）を含む合成文を埋め込み，PCA により 2 次元に可視化した（図 1）．視覚的にも，PC1 方向に沿って両群が明瞭に分離する．

定量的には，PC1 の分散説明率は EVR(PC1) ≈ 0.105，ラベル（ego/allo）との関係は $\eta^2 \approx 0.6$ と大きく，群平均もほぼ対称（ego が正，allo が負）に割れていた．さらに，方向キー（E/W/S/N vs F/B/L/R）別の PC1 平均においても，allocentric 側はおおかた負，egocentric 側は正となり，「PC1 の両端に寄る」性質が直接確認できる．バイオリン図では，つぎのようになる．図 2 が合成文，図 3 が，ウィキペディアから任意に取った自然文からである．

加えて，プローブベクトル v_{probe} と PC1 方向 u_1 の整列度は高く（例：$|\cos(v_{\text{probe}}, u_1)| \approx 0.60$–$0.66$），PCA の PC1 が「ego 語群 ↔allo 語群」の差分方向を拾っているという解釈を支持する．

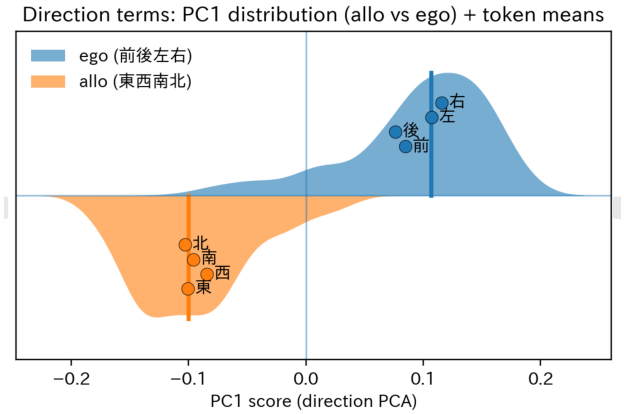

図 3　方向語を含む自然文ベクトルの PCA の PC1 での分布図（拡大したものを付録ページに掲載）

本研究の生成データでは，語形レベルのプローブ（例：前/後/左/右 → 東/西/南/北 および 前方/後方/左方/右方 → 東方/西方/南方/北方）に対して align ≈ 0.66（hou）および ≈ 0.60（kanji1）と高い整列が得られ，句レベル（前後左右 → 東西南北）でも align ≈ 0.37 を示した．これは，埋め込み空間内に「ego 語群 ↔ allo 語群」を結ぶ差分方向が明確に立っており，PCA の PC1 がそれを拾ったという解釈を支持する．

5.2　交絡チェック：文長・句読点・テンプレート差

合成文生成では，文長・句読点・スロット位置・テンプレート ID が方向語群間で偏らないよう設計した．さらに，文長や句読点等を共変量として PC1 スコアを残差化した場合でも，ego/allo の分離（η^2）は低下せず，少なくとも「長さや句読点の差が PC1 を作った」という型の疑似相関では説明しにくい．

5.3　2×2（frame cue × token）：主効果と相互作用

2×2 要因計画に基づき，(i) 語彙主効果 Δ_{token} と (ii) 相互関係 v_{int} を比較した．方向語では，Δ_{token} が支配的で v_{int} は相対的に小さく，文脈 cue に依存した局所的増幅ではなく「語彙対立そのものが global な軸を作る」構造が支持された．

5.4　背景主成分の診断・除去（小説背景／地理背景）

方向語 PC1 が「背景コーパスの支配的主成分に寄生した疑似相関」である可能性を検証するため，背景コーパスに対する PCA（背景 PC）を用いた診

表 1　背景 PC 除去（rm 条件）に対する方向語分離の頑健性（Wikipedia 地理背景）．η^2 は ego/allo ラベルが PC1 スコア分散を説明する割合（ANOVA）．sep は群平均との差（ego−allo）の大きさ．

removed bgPCs	η^2	retention	sep（ego−allo）
0	0.679	1.000	0.178
1	0.673	0.991	0.177
5	0.582	0.858	0.158
10	0.582	0.858	0.158

断と除去を行った．背景として (a) 小説文（明治文豪の複数作家）と (b) Wikipedia 由来の地理・地形記述（空間語彙が密なストレステスト）を用い，rm1（PC1 除去，以下同じ）/rm5/rm10 および，軸を増加した $K = 50$ など複数条件で背景 PC を除去した．

結果として，いずれの背景においても，支配的背景 PC を除去しても方向語の PC1 分離は崩れにくかった．地理背景では，背景 PC のうち PC3/PC4 等との部分整列が観察される場合があったが，これは空間セマンティクス領域の共有として自然に説明できる可能性があり，少なくとも「背景 PC1 への単純な寄生」だけでは説明しにくい（表 1）

5.5　方向語軸へのアンカー射影：deixis の偏在

方向語が作る global な二極軸と，他のダイクティック表現が示す局所差分を同一座標系で比較するため，方向語データのみで推定した direction-PC1 をアンカーとして固定し，他データをその軸へ射影した（section 4.4）．対象は，(i) 〜ていく／〜てくる（YK），(ii) 授受（〜てやる／〜てくれる／〜てもらう；YMO），(iii) こ／そ（こそ;KOSO）である．

射影結果図（図 4）では，方向語は 0 をまたいで両側に分布する（bipolar/global）のに対し，YK/YMO/KOSO は主として egocentric 側半空間に偏る（polarization）傾向を示した．すなわち，方向語の参照枠対立が「空間全体を二分する global な軸」として現れやすい一方，これらの deixis は「egocentric 領域内での局所差分」として現れている可能性が高い．

6　考察

6.1　方向語 PC1 は単純な文体差・背景寄生では説明しにくい

合成文生成による交絡統制（文長・句読点・テンプレート）に加え，背景 PC の診断・除去後も方向語分離が維持されたことから，少なくとも本研究の範囲では，方向語 PC1 は単純な文体差やデータセット固有の支配的主成分への寄生のみでは説明しにくい．特に地理背景は語彙領域が近く，疑似相関が生じやすい条件であるが，その条件下でも分離が崩れにくい点は重要である．

6.2　egocentric-allocentric axis（二極）と polarization（極性化）

方向語は PC1 に沿った global な二極分離として現れやすい一方で，YK/YMO/KOSO は direction-PC1 への射影で非対称性を示した．この対比は，(i) 空間参照枠のように全体空間を二分する軸（egocentric–allocentric）と，(ii) egocentric 側で主に顕在化する局所差分（deixis ／ egophoricity 的現象）の階層的関係を示唆する．この見取り図は，**言語学的直観（指示・授受・移動が話し手側の領域で活発になる）とも整合的**である．（図 4 参照）

7　結論

本研究は，日本語方向語を手がかりとして，文埋め込み空間に空間認知表現が線形に抽出可能な形で潜在しているかを検討した．合成文埋め込みでは，PC1 に沿った明瞭な分離が観察され，方向キー別平均やプローブベクトル整列からも PC1 の解釈（ego/allo 軸）が支持された．さらに，背景主成分の診断・除去（小説背景／地理背景）後も分離が崩れにくく，単純な文体差や背景寄生のみでは説明しにくい頑健性が示唆された．加えて，方向語で定義した軸へのアンカー射影により，「てゆく・てくる」指示詞等の deixis は，egocentric 側半空間で極性的に現れ，空間参照枠で代表される egocentric-allocentric axis（二極軸）と egocentric 内部差分の階層構造が示唆された．今後は，複数モデル比較や他の egophoric な表現・他言語への拡張により，この表現幾何の一般性を検討する予定である．

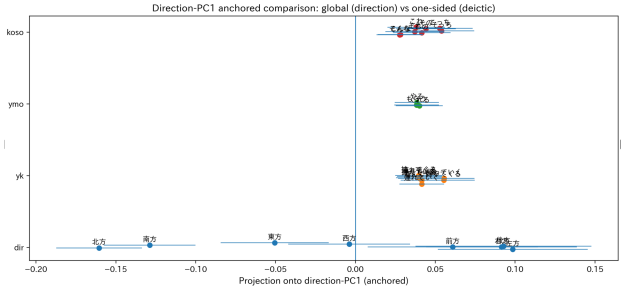

図 4　PC1 への方向語とダイクシスの射影図（拡大したものを付録ページに掲載）

謝辞

本研究にあたっては，国立国語研究所の「書き言葉均衡コーパス (BCCWJ)」のデータを利用した部分がある．感謝申し上げたい．

参考文献

[1] Stephen C. Levinson. **Space in Language and Cognition: Explorations in Cognitive Diversity**. Cambridge University Press, Cambridge, 2003.

[2] Patrick Byrne, Suzanna Becker, and Neil Burgess. Remembering the past and imagining the future: a neural model of spatial memory and imagery. **Psychological Review**, Vol. 114, No. 2, pp. 340–375, 2007.

[3] 渡辺実. 「わがこと・ひとごと」の観点と文法論. 『国語学』, No. 165, pp. 1–15, 1991.

[4] 近藤 泰弘・澤田淳. 日本語文法研究の射程. 開拓社, Tokyo, 2025.

[5] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. **arXiv preprint arXiv:2212.03533**, 2022.

[6] Adnan Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsay, Tamera Bricken, Brian Chen, Adam Pearce, Nicholas Ley, Hatfield Russell, Tom Henighan, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. **Transformer Circuits Thread**, 2024.

[7] Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel B. M. Haun, and Stephen C. Levinson. Can language restructure cognition? the case for space. **Trends in Cognitive Sciences**, Vol. 8, No. 3, pp. 108–114, 2004.

[8] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, 2019.

[9] Wes Gurnee and Max Tegmark. Language models represent space and time. **Proceedings of the 41st International Conference on Machine Learning (ICML)**, Vol. 235, , 2024.

[10] Cunningham Bricken, Adnan Templeton, Josiah Braun, Brian Jiao, Joshua Batson, Tom Conerly, Catherine Olsson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. **Transformer Circuits Thread**, 2023.

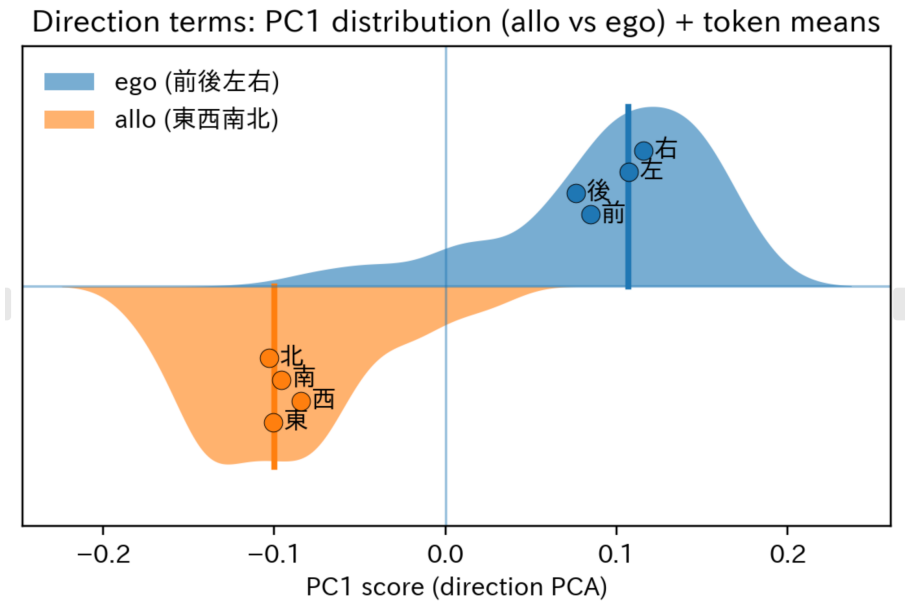

図 5　【付録 (Appendix)】方向語を含む自然文ベクトルの PCA の PC1 での分布図 (図 3 の拡大図)

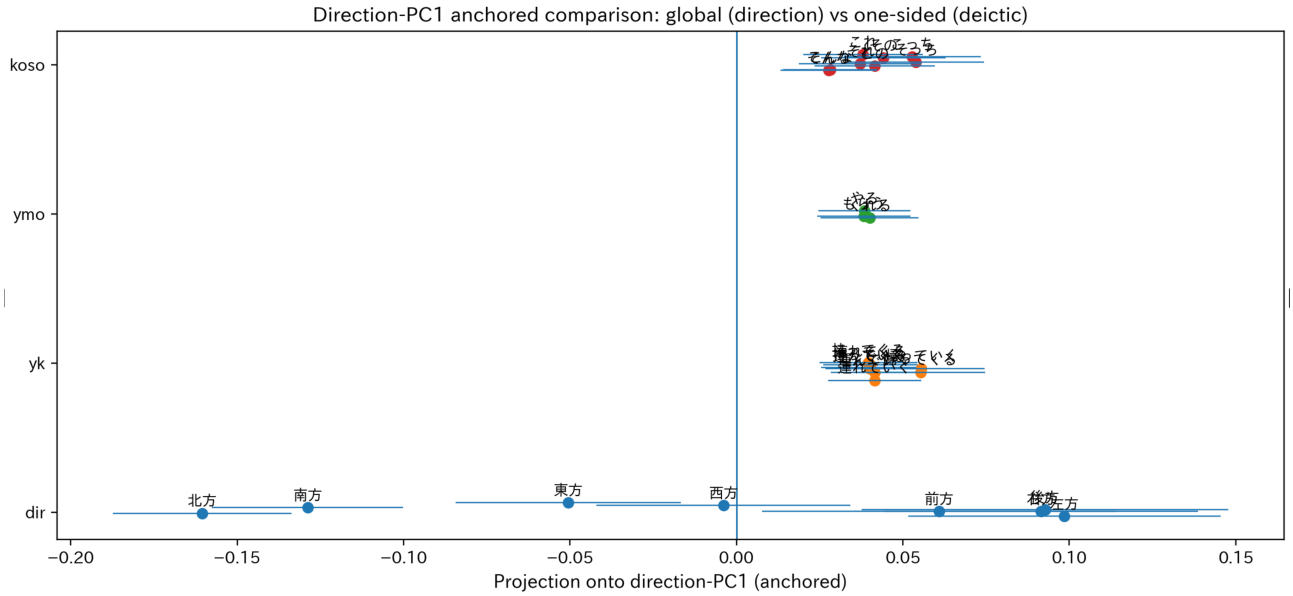

図 6　【付録 (Appendix)】PC1 への方向語とダイクシスの射影図 (図 4 の拡大図・下方が方向語．上方が順に，指示詞・やりもらい・「ていく・てくる」である．上下は単なる見た目のオフセットである．