Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Author Name Disambiguation in Portuguese Scientific Publications: The SBC-AND dataset

Natan de S. Rodrigues

natan.rodrigues@ueg.br

State University of Goiás https://orcid.org/0000-0002-0785-4397

Samuel Gomes dos Santos

State University of Goiás

Vitória Maria Diniz

State University of Goiás

Sirlene A. Rodrigues Costa

State University of Goiás

Research Article

Keywords: AND, SBC-AND dataset, bibliographic repositories, Portuguese language, scientific metadata

Posted Date: July 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10284834/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

DESAMBIGUAÇÃO DE NOMES DE AUTORES EM PUBLICAÇÕES CIENTÍFICAS EM PORTUGUÊS: O *DATASET* SBC-AND

AUTHOR NAME DISAMBIGUATION IN PORTUGUESE SCIENTIFIC PUBLICATIONS: THE SBC-AND DATASET

Natan de S. Rodrigues (UEG)[1]

Samuel Gomes dos Santos (UEG)[2]

Vitória Maria Diniz (UEG)[3]

Sirlene A. Rodrigues Costa (UEG)[4]

Resumo: A desambiguação de nomes de autores (AND) contribui para a qualidade dos dados em repositórios bibliográficos. Apesar dos avanços recentes, a maioria dos conjuntos de dados (*datasets)* públicos para AND concentra-se em publicações em inglês, deixando o português pouco representado. Este estudo apresenta o SBC-AND, um dataset curado e padronizado para AND em publicações científicas em português. O dataset reúne 442 registros de periódicos da Sociedade Brasileira de Computação, organizados em 21 blocos ambíguos e 232 *clusters* reais de autores. Os registros incluem título, autores, coautores, resumo, palavras-chave, veículo e ano de publicação. O SBC-AND foi avaliado com um *framework* configurável para AND, considerando *embeddings* multilíngues, camadas de Redes Convolucionais em Grafo e épocas de treinamento. Os resultados indicam que o *dataset* é um conjunto de referência (*benchmark)* desafiador, especialmente em blocos com sobrenomes frequentes e baixa recorrência de registros por autor. O SBC-AND está disponível em um repositório *online* para apoiar a reprodutibilidade e novas pesquisas em AND.

Palavras-chave: AND; dataset SBC-AND; repositórios bibliográficos; língua portuguesa; metadados científicos.

Abstract: *Author Name Disambiguation (AND) contributes to data quality in bibliographic repositories. Despite recent advances, most public AND datasets focus on English publications, leaving Portuguese underrepresented. This study presents SBC-AND, a curated and standardized dataset for AND in Portuguese scientific publications. The dataset contains 442 records from journals of the Brazilian Computer Society, organized into 21 ambiguous blocks and 232 real author clusters. The records include title, authors, coauthors, abstract, keywords, venue, and publication year. SBC-AND was evaluated using a configurable framework for AND, considering multilingual embeddings, Graph Convolutional Network layers, and training epochs. The results indicate that the dataset is a demanding benchmark, especially in blocks with frequent surnames and low record recurrence per author. SBC-AND is available in an online repository to support reproducibility and future research on AND.*

Keywords*: AND; SBC-AND dataset; bibliographic repositories; Portuguese language; scientific metadata.*

---

[1] Docente na Universidade Estadual de Goiás (UEG). Doutor em Informática pela Universidade de Brasília (UnB). E-mail: natan.rodrigues@ueg.br

[2] Discente do Curso Superior de Tecnologia em Sistemas para Internet da Universidade Estadual de Goiás (UEG). E-mail: samuel.santos@aluno.ueg.br.

[3] Discente do Curso Superior de Tecnologia em Sistemas para Internet da Universidade Estadual de Goiás (UEG). E-mail: vivitoria111@aluno.ueg.br.

[4] Docente na Universidade Estadual de Goiás (UEG). Doutora em Estudos Linguísticos pela Universidade Federal de Minas Gerais (UFMG). E-mail: sirleneletras@bol.com.br

1 Introduction

Identifying authors correctly in scientific publications is essential to maintain the accuracy of bibliographic data. When the same name refers to different researchers, or when one author appears with multiple name forms, the information in databases becomes unreliable. This challenge, known as Author Name Disambiguation (AND), influences citation counts, collaboration networks, and many indicators used to evaluate research (Ferreira; Gonçalves; Laender, 2020).

Studies on this topic have evolved for more than twenty years (Rodrigues; Mariano; Ralha, 2024). Early work compared author names, titles, and coauthors using rules and string similarity (Cota et al., 2010). Later research began to apply clustering and learning methods that combined several metadata attributes to improve precision (Ferreira et al., 2014). More recent approaches use text embeddings and graph models to capture both the meaning and the relationships among publications (Zhou et al.,2024; Rodrigues; Ralha, 2026). These strategies improved results in international datasets such as DBLP[5] and AMiner[6]. Even so, most of these advances focus on English data, while publications in other languages receive little attention.

Repositories that store Portuguese publications face specific difficulties. Many records lack author identifiers or present names written in inconsistent ways, with abbreviations, missing accents, or truncations. A study of eight institutional repositories in Portugal found that more than ninety percent of author entries had no ORCID[7] or CIÊNCIA ID[8], and about 15% contained inconsistencies in the name field (Rodrigues; Rodrigues, 2024). The authors of that study pointed out that the absence of authority control and the linguistic complexity of Portuguese names make AND tasks harder. They also stressed the need for standardized identifiers and better metadata quality to improve author identification.

Portuguese publications face similar problems. Author names often include multiple surnames or particles such as "*de*", "*do*", "*da*", "*dos*" and "*das*", which appear differently across venues (Rodrigues et al., 2020). The BDBComp dataset[9] (Cota et al., 2010) was one of the first initiatives to explore this issue in Computer Science (CS) publications in

[5] https://dblp.org/
[6] https://www.aminer.cn/
[7] https://orcid.org/
[8] https://www.ciencia-id.pt
[9] https://github.com/kashak79/DISC_AND_Data_Set/blob/master/bdbcomp.tar

Portuguese. It supported early heuristic and clustering methods like SAND (Ferreira et al., 2014), but its format is not standardized for current frameworks. Without consistent data organization, researchers find it difficult to compare results or reproduce experiments reliably.

Recent systems for AND adopt flexible designs that adapt to different types of data. They combine embeddings, graph representations, and clustering techniques while allowing configuration of parameters such as similarity thresholds or graph depth (Zhou et al., 2024; Rodrigues; Ralha, 2025; Rodrigues; Ralha, 2026). These systems make it possible to test AND tasks in other languages and datasets that vary in structure and completeness.

This study presents SBC-AND, a curated and standardized dataset of Portuguese scientific publications from journals of the Brazilian Computer Society (SBC). The dataset groups ambiguous author names, includes full metadata, and follows a JSON[10] format compatible with configurable AND frameworks (Rodrigues; Ralha, 2025; Rodrigues; Ralha, 2026). Using a flexible and configurable framework for AND, we test how current methods perform when applied to Portuguese data and analyze the results in detail.

This work contributes in three main ways:

- It introduces SBC-AND, a new dataset for AND in Portuguese with standardized and complete metadata.
- It uses a configurable framework for AND to evaluate the proposed dataset under different experimental configurations.
- It presents quantitative and qualitative analyses that characterize the challenges of AND in Portuguese bibliographic records and support future studies on multilingual and low-resource bibliographic data.

The next sections are organized as follows. Section 2 reviews related research and datasets. Section 3 describes the SBC-AND dataset and its curation. Section 4 explains the experimental setup and the system configuration. Section 5 presents and discusses the results. Section 6 concludes with final remarks and future directions.

[10] https://www.ecma-international.org/publications-and-standards/standards/ecma-404/

2 Related Work

Benchmark datasets have played a central role in advancing research on AND. Among the most widely used resources, the AMiner collections provide large-scale bibliographic data for evaluating AND methods.

Tang et al. (2012) released one of the earliest publicly documented datasets for AND, based on records extracted from AMiner. Zhang et al. (2018) later extended this line of work with AMiner-18, combining metric and graph-based representation learning with human validation of author clusters. Other datasets rely on DBLP metadata, such as the subsets used by Qian et al. (2015) to evaluate BatchAD and IncAD in dynamic digital libraries, and WhoIs-DBLP, proposed by Boukhers and Asundi (2022), which provides ambiguous author blocks for testing deep learning models with open and reproducible data. Larger benchmarks have also been proposed. S2AND integrates data from Semantic Scholar, ORCID, Microsoft Academic Graph, and DBLP, offering labeled clusters and standardized splits for supervised and unsupervised models (Subramanian et al., 2021). LAGOS-AND links Microsoft Academic Graph and ORCID identifiers to provide a large gold-standard dataset with block-level and pairwise annotations for clustering and classification evaluation (Zhang et al., 2023).

These datasets have improved reproducibility and comparability in AND. However, they rely mainly on English metadata and do not capture linguistic patterns typical of Portuguese names. BDBComp is the only previous effort focused on Brazilian publications (Cota et al., 2010), supporting heuristic and clustering methods such as SAND (Ferreira et al., 2014), but its structure is not standardized for integration with modern frameworks such as ADAN (Rodrigues; Ralha, 2025; Rodrigues; Ralha, 2026). SBC-AND addresses this gap by providing curated and structured Portuguese records for reproducible evaluation in an underrepresented linguistic context.

Table 1 summarizes the main public datasets used in AND research. It lists the number of documents, author blocks, and true authors when available, along with the main data source. This comparison highlights how existing benchmarks have evolved in scale and standardization and how SBC-AND complements them with data in Portuguese.

Table 1 - Comparison of public datasets for AND.

Dataset	Documents	Author Blocks	True Authors	Main Source
AMiner-12 (Tang et al., 2012)	7447	109	1546	ArnetMiner system
AMiner-18 (Zhang et al., 2018)	70,258	100	12,798	AMiner platform
DBLP (Qian et al., 2015)	6478	679	1463	DBLP repository
S2AND (Subramanian et al., 2021)	487,567	27,628	59,119	SS + ORCID + MAG
LAGOS-AND (Zhang et al., 2023)	7.5M	798	798	MAG + ORCID
WhoIs-DBLP (Boukhers et al., 2022)	--	3,000+	3,000+	DBLP + Zenodo
BDBComp (Cota et al., 2010)	3,000+	30	130	Brazilian CS journals
SBC-AND (ours)	442	21	232	SBC journals (Portuguese)

3 The SBC-AND Dataset

The SBC-AND dataset was developed to support research on AND in Portuguese-language scientific publications. It compiles metadata from open-access journals maintained by the SBC to capture ambiguities that are frequent in Portuguese author names. The dataset includes articles from *Journal of the Brazilian Computer Society*, *Journal of Information and Data Management*, *Journal on Interactive Systems*, *Journal of Software Engineering Research and Development*, *International Journal of Computer Architecture Education*, and *Brazilian Journal of Information Systems*. All journals are hosted on the SBC's open-access platform SOL[11]. All journals are hosted on SBC's open-access SOL platform. A Python script collected metadata from each article page, including authors, title, abstract, coauthors, keywords, journal, and publication year. Only complete research papers with sufficient metadata were included. The dataset focuses on publications by Brazilian or Portuguese researchers to capture linguistic patterns typical of the Lusophone context.

After collection, manual curation was performed to identify real authors and resolve name ambiguities. The records were grouped into ambiguous author blocks based on shared abbreviated signatures, such as "*A. Silva*" or "*A. Araujo*". Curators inspected each block using evidence such as title, affiliation, coauthors, research area, and full name across publications. This process produced verified real author clusters used as reference data for evaluating AND models.

Each ambiguous block appears as an individual JSON file named after the corresponding author signature, for example *a_araujo.json*. Each file groups all articles linked

[11] https://sol.sbc.org.br

to that abbreviated name and stores fields such as title, authors, venue, year, abstract, and keywords. Figure 1 shows a partial example of this structure with a single article for illustration.

Figure 1 - Partial example of a JSON file for an ambiguous author block in SBC-AND.


```
[
  {
    "id": "0",
    "label": "1",
    "author": "a araujo",
    "title": "A produção de material didático para a EaD na perspectiva dos princípios dialógicos",
    "abstract": "A educação a distância avança e contribui cada vez mais com o ensino e a
        aprendizagem, favorecendo a criação de estratégias que aproximam o professor autor e o
        estudante, apoiados pelas tecnologias digitais da informação e comunicação...",
    "coauthors": [
      "e santos"
    ],
    "keywords": [
      "Material didático",
      "Educação a distância",
      "Princípios dialógicos"
    ],
    "venue": "Revista Brasileira de Informática na Educação",
    "year": "2025"
  }
]
```


Source: Created by the authors

3.1 Quantitative Profile of the SBC-AND Dataset

The structure of SBC-AND relies on two grouping levels. The first level consists of external blocks, which group records with the same abbreviated author signature, such as *a_silva* or *m_santos*. The second level consists of real clusters within each block. Each real cluster represents one verified author. This structure allows two types of analysis. The global analysis describes the dataset as a whole, while the block-level analysis describes the internal composition of each ambiguous signature.

Let $N_{records}$ denote the total number of publication records, N_{blocks} the number of external blocks, and $N_{authors}$ the number of real clusters in the dataset. Equations 1-3 describe global averages that summarize overall density and author recurrence:

1. $Avg^{global}_{records/block} = \frac{N_{records}}{N_{blocks}}$
2. $Avg^{global}_{authors/block} = \frac{N_{authors}}{N_{blocks}}$
3. $Avg^{global}_{records/author} = \frac{N_{records}}{N_{authors}}$

The Equation 1 indicates the average size of the ambiguous-name blocks. The Equation 2 describes how many distinct verified authors are represented, on average, within each external block. The Equation 3 captures the average recurrence of each verified author in the dataset.

SBC-AND contains 442 publication records distributed across $N_{blocks} = 21$ external blocks and $N_{authors} = 232$ real clusters. The dataset shows $Avg^{global}_{records/block} = 21.05$,

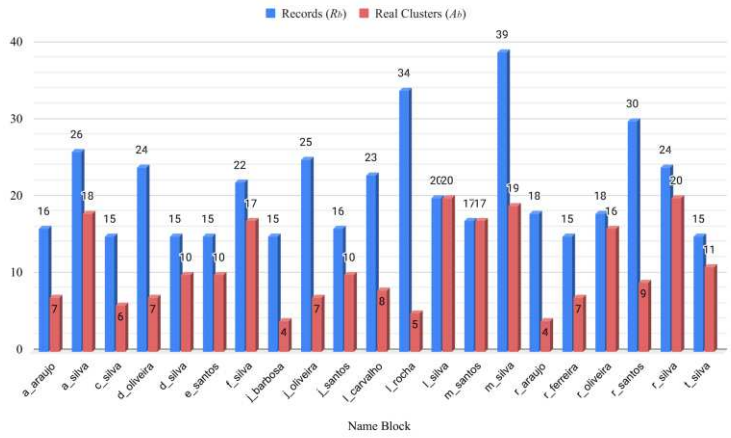
$Avg^{global}_{authors/block} = 11.05$, and $Avg^{global}_{records/author} = 1.90$. These values summarize the size and basic structure of the collection and describe how records and real clusters distribute across the external blocks.

At the block scale, each external block b contains R_b records and A_b real clusters. Equation 4 defines the average number of records per author inside a specific block:

4. $Avg_{records/author}(b) = \frac{R_b}{A_b}$

This ratio captures how much evidence each real author contributes inside one ambiguous signature. Blocks with low $Avg_{records/author}(b)$ present many clusters with only one record. Blocks with higher values show authors with multiple publications but still share space with many unrelated individuals.

Figure 2 - Records (R_b) and real clusters (A_b) in each external block of the SBC-AND.



Source: Created by the authors

Figure 2 reports R_b and A_b for every external block. Groups such as *l_silva*, *m_silva*, and *r_silva* combine high values of R_b with many real clusters. This pattern follows the frequency of Portuguese surnames in Brazil, where a small set of family names covers a large portion of the population (Monasterio, 2016). These traits increase the probability that unrelated authors share identical abbreviated signatures. The distribution in Figure 2 shows that external blocks vary widely in size, and real clusters display heterogeneity. This structure creates an environment for evaluating AND methods because frequent surname collision and low record recurrence reduce the separation between identity profiles.

4 Experimental Evaluation

This section presents the experimental evaluation of SBC-AND. It describes the validation metrics, experimental setup, embedding models, graph-based configurations, and training parameters used in the experiments.

4.1 Validation Metrics

According to (Ferreira; Gonçalves; Laender, 2020), Pairwise Precision (pP), Pairwise Recall (pR), Pairwise F1 ($pF1$), Average Cluster Purity (ACP), Average Author Purity (AAP), and K-metric (K) are commonly used metrics for evaluating AND methods. Following this evaluation practice, this work adopts these metrics to assess the AND results on SBC-AND. Pairwise metrics evaluate whether pairs of records were correctly grouped or separated, while ACP, AAP, and K evaluate the overall structure of the generated clusters.

Let TP denote the number of pairs of records that belong to the same real author and are also assigned to the same predicted cluster. Let FP denote the number of pairs assigned to the same predicted cluster but belonging to different real authors. Let FN denote the number of pairs that belong to the same real author but are assigned to different predicted clusters. The pP, shown in Equation 5, measures the proportion of predicted author links that are correct.

5. $$pP = \frac{TP}{(TP+FP)}$$

The metric pR, shown in Equation 6, measures the proportion of real author links that were recovered by the predicted clustering.

6. $$pR = \frac{TP}{(TP+FN)}$$

$pF1$, shown in Equation 7, is the harmonic mean between pP and pR, providing a single score that balances both criteria.

7. $$pF1 = \frac{2 \cdot (pP \cdot pR)}{(pP + pR)}$$

In addition, we report ACP, AAP, and K. Let N be the total number of records, e the number of predicted clusters, t the number of real author clusters, n_{ij} the number of records shared by predicted cluster i and real cluster j, n_i the number of records in predicted cluster i, and n_j the number of records in real cluster j. The ACP, shown in Equation 8, evaluates the

purity of each generated cluster, that is, the extent to which its records belong to the same real author.

$$8.\quad ACP = \frac{1}{N}\sum_{i=1}^{e} \quad \sum_{j=1}^{t} \quad \frac{n_{ij}^2}{n_i}$$

The AAP metric, shown in Equation 9, evaluates the opposite perspective by measuring how concentrated each real author cluster is within the predicted clusters.

$$9.\quad AAP = \frac{1}{N}\sum_{j=1}^{t} \quad \sum_{i=1}^{e} \quad \frac{n_{ij}^2}{n_i}$$

Finally, the K, shown in Equation 10, is defined as the harmonic mean between ACP and AAP.

$$10.\ K = \sqrt{ACP \cdot AAP}$$

Together, these metrics provide a complete evaluation of the AND task results. The $pF1$ assesses local pairwise decisions, whereas the K captures the broader clustering structure produced by the adopted AND method.

4.2 Experimental Setup

The experiments were conducted using ADAN[12] (Rodrigues; Ralha, 2025; Rodrigues; Ralha, 2026), a configurable framework for AND that integrates textual embeddings, heterogeneous graph construction, representation learning with Graph Convolutional Networks (GCN), clustering, and evaluation in a unified pipeline. Each ambiguous author block from SBC-AND was processed independently, producing one clustering solution per block. All experiments were executed on a *MacBook Air* with an *Apple M1* processor, 16 GB of RAM, a 256 GB SSD, and *macOS Sequoia*.

We evaluated two multilingual embedding models: *BAAI/bge-m3*[13] (Chen et al., 2024) and *ibm-granite/granite-embedding-311m-multilingual-r2*[14], referred to here as *IBM Granite* (Awasthy et al., 2026). These models were used to represent textual metadata from each publication record, including titles, abstracts, and keywords when available. For each embedding model, we tested nine graph-based configurations by varying the number of GCN

[12] https://github.com/natansr/adan

[13] https://huggingface.co/BAAI/bge-m3

[14] https://huggingface.co/ibm-granite/granite-embedding-311m-multilingual-r2

layers and training epochs. We used one (1), two (2), or three (3) GCN layers and 1000, 2000, or 3000 epochs, resulting in 18 configurations in total.

In each configuration, ADAN builds a heterogeneous graph for each ambiguous block, with publication records as nodes and bibliographic evidence, such as coauthorship and venues, as graph relations. The embeddings are used as node features, and the GCN learns graph-aware representations before clustering. The resulting clusters represent predicted author identities. All configurations used the same blocks and evaluation protocol, allowing direct comparison of the embedding model, GCN depth, and training duration.

5 Results and Analysis

This section presents the results obtained after the experimental execution and discusses the performance of ADAN on SBC-AND from both a global and a block-level perspective.

Table 2 summarizes the results for all evaluated configurations. The first clear pattern is that SBC-AND is difficult for AND tasks when the evaluation focuses on Pairwise decisions. The $pF1$ scores are much lower than the values obtained for ACP, AAP, and K. This means that ADAN often produces clusters that are reasonably consistent with the true author groups, but still has difficulty recovering all records that belong to the same real author. *IBM Granite* achieved the best results for the Pairwise metrics. Its average $pF1$ was 0.4497, compared with 0.3640 for *BAAI/bge-m3*. The best $pF1$ score was 0.4600, obtained with *IBM Granite*, two GCN layers, and 1000 epochs. This suggests that, in this dataset, *IBM Granite* produced textual representations that were more useful for separating author identities.

Table 2 - Experimental results for all ADAN configurations on SBC-AND.

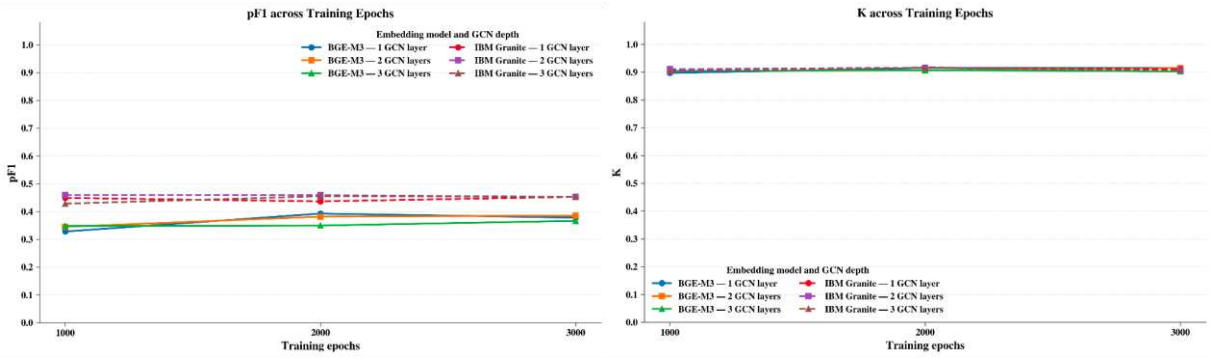
Embedding Model	GCN Layers	Epochs	pP	pR	pF1	ACP	AAP	K
BAAI/bge_m3	1	1000	0.3806	0.3050	0.3282	0.9011	0.8936	0.8970
BAAI/bge_m3	1	2000	0.4341	0.3703	0.3928	0.9251	0.9065	0.9155
BAAI/bge_m3	1	3000	0.4341	0.3438	0.3780	0.9206	0.9098	0.9149
BAAI/bge_m3	2	1000	0.3740	0.3298	0.3453	0.9090	0.9005	0.9045
BAAI/bge_m3	2	2000	0.4241	0.3595	0.3818	0.9082	0.9060	0.9067
BAAI/bge_m3	2	3000	0.4120	0.3750	0.3854	0.9233	0.9066	0.9147
BAAI/bge_m3	3	1000	0.3618	0.3444	0.3479	0.9128	0.8955	0.9038

BAAI/bge_m3	3	2000	0.3813	0.3348	0.3498	0.9121	0.9033	0.9073
BAAI/bge_m3	3	3000	0.4159	0.3374	0.3666	0.9085	0.8982	0.9030
IBM Granite	1	1000	0.4962	0.4207	0.4491	0.9011	0.9129	0.9067
IBM Granite	1	2000	0.4565	0.4341	0.4367	0.9207	0.9131	**0.9168**
IBM Granite	1	3000	0.4887	0.4374	0.4530	0.9082	0.9145	0.9112
IBM Granite	2	1000	0.4910	0.4430	**0.4600**	0.9089	0.9134	0.9109
IBM Granite	2	2000	0.4833	**0.4508**	0.4591	**0.9168**	0.9157	0.9161
IBM Granite	2	3000	**0.5046**	0.4280	0.4529	0.9020	0.9111	0.9063
IBM Granite	3	1000	0.4663	0.4074	0.4284	0.8985	0.9084	0.9032
IBM Granite	3	2000	0.4924	0.4406	0.4552	0.9132	**0.9183**	0.9155
IBM Granite	3	3000	0.4985	0.4316	0.4532	0.9041	0.9099	0.9067

ACP, AAP, and K remained high in almost all configurations. K was always above 0.89 and reached its best value, 0.9168, with *IBM Granite*, one GCN layer, and 2000 epochs. ACP and AAP showed a similar trend. This difference between $pF1$ and K metrics is expected, since they capture different aspects of the clustering result. Pairwise F1 is more sensitive to local pairwise decisions, whereas ACP, AAP, and K describe the overall agreement between the generated clusters and the real author groups. Thus, ADAN may preserve a coherent global clustering structure even when some individual record associations are not recovered.

Figure 3 shows pF1 and K across training epochs, embedding models, and GCN layers. *IBM Granite* outperformed *BAAI/bge-m3* in $pF1$ in all configurations, with the best results obtained using two GCN layers. Three layers did not improve performance, which may indicate that deeper propagation introduces noise or excessive smoothing in the learned representations. The metric K, however, remained high and more stable, with smaller differences across embeddings and GCN layers. This suggests that ADAN maintained globally consistent clusters, even when pairwise recovery varied across configurations.

Figure 3 - $pF1$ and K scores across embedding models, GCN layers, and training epochs.

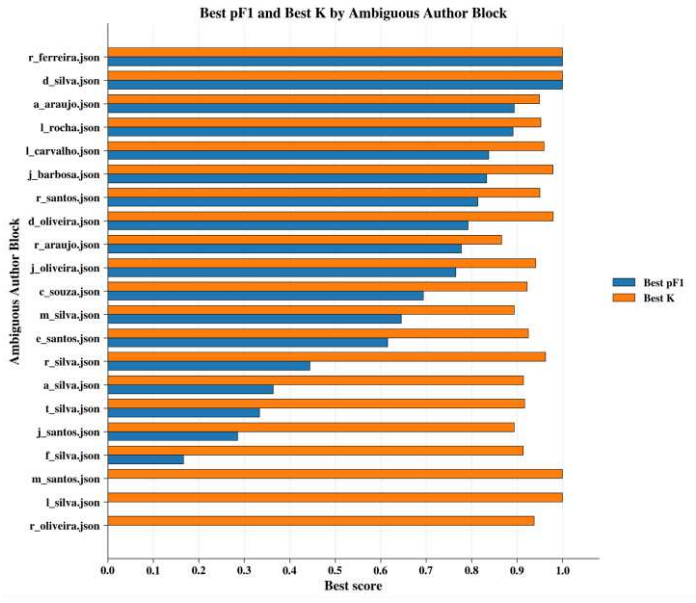


Source: Created by the authors

Figure 4 shows the best pF1 and K values for each ambiguous author block. The results show that difficulty varies across SBC-AND. Blocks such as r_oliveira, l_silva, and m_santos reached a best pF1 of 0.00, although their K values remained high. This occurs because these blocks include many real authors represented by only one or very few records, reducing the positive evidence captured by Pairwise metrics.

Other blocks, such as *f_silva*, *j_santos*, *t_silva*, *a_silva*, and *r_silva*, also remained difficult, with best $pF1$ values below 0.45. These cases reflect realistic ambiguity scenarios, combining common abbreviated signatures with low author recurrence. In contrast, blocks such as *l_rocha*, *a_araujo*, *d_silva*, and *r_ferreira* achieved higher $pF1$ values, suggesting that ADAN performs better when real authors have more repeated records.

Figure 4 - Best $pF1$ and K scores obtained for each ambiguous author block in SBC-AND.



Source: Created by the authors

Overall, these results show that SBC-AND is a demanding benchmark for AND. The dataset includes cases in which the global clustering structure remains consistent, while pairwise identity recovery is still limited. This makes SBC-AND useful for evaluating AND methods under realistic conditions, especially when the analysis goes beyond global averages.

6 Conclusion and Future Work

This study introduced SBC-AND, a curated and standardized dataset for AND in Portuguese scientific publications. The dataset provides ambiguous author blocks, verified real author clusters, and structured metadata in an online repository to support reuse and reproducibility[15]. The evaluation showed that SBC-AND is a demanding benchmark, with high ACP, AAP, and K values, but lower pairwise results in blocks with common surnames and low record recurrence. *IBM Granite* achieved the best overall pairwise performance, especially with two GCN layers, while deeper graph propagation did not improve the results. Future work includes expanding SBC-AND with more journals and conference proceedings, adding metadata and author identifiers when available, and evaluating other AND methods for broader comparisons in Portuguese and multilingual data.

REFERENCES

AWASTHY, Parul; TRIVEDI, Aashka; YANG, Yushu; BARKER, Ken; LI, Yulong; IYER, Bhavani; FRANZ, Martin; BROSS, Juergen; DOSHI, Meet; P, Vignesh; KUMAR, Vishwajeet; WARD, Todd; DANIELS, Abraham; LEE, Madison; LASTRAS, Luis; SEN, Jaydeep; FLORIAN, Radu. **Granite Embedding Multilingual R2 Models.** arXiv, 2026. Available at: https://arxiv.org/abs/2605.13521. Accessed on: 8 Jun. 2026.

BOUKHERS, Zeyd; ASUNDI, Nagaraj Bahubali. **Whois? Deep author name disambiguation using bibliographic data.** In: Linking Theory and Practice of Digital Libraries. Cham: Springer International Publishing, 2022. p. 201–215.

CHEN, Jianlv; XIAO, Shitao; ZHANG, Peitian; LUO, Kun; LIAN, Defu; LIU, Zheng. **BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation.** arXiv, 2024. Available at: https://arxiv.org/abs/2402.03216. Accessed on: 8 Jun. 2026.

[15] https://github.com/natansr/sbc-and

COTA, Ricardo G.; FERREIRA, Anderson A.; NASCIMENTO, Cristiano; GONÇALVES, Marcos André; LAENDER, Alberto H. **An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations.** Journal of the American Society for Information Science and Technology, v. 61, n. 9, p. 1853–1870, 2010.

FERREIRA, A. A.; GONÇALVES, M. A.; LAENDER, A. H. F. **A brief survey of automatic methods for author name disambiguation.** ACM SIGMOD Record, v. 41, n. 2, p. 15–26, 2012.

FERREIRA, Anderson A.; GONÇALVES, Marcos André; LAENDER, Alberto H. F. **Automatic Disambiguation of Author Names in Bibliographic Repositories.** San Rafael: Morgan & Claypool Publishers, 2020.

FERREIRA, Anderson A.; VELOSO, Adriano; GONÇALVES, Marcos André; LAENDER, Alberto H. F. **Self-training author name disambiguation for information scarce scenarios.** Journal of the Association for Information Science and Technology, v. 65, n. 6, p. 1257–1278, 2014.

MONASTERIO, Leonardo Monteiro. **Sobrenomes e ancestralidade no Brasil.** 2016. Available at: https://repositorio.ipea.gov.br/server/api/core/bitstreams/c12836fd-73a7-4248-bd45-74f6acff772a/content. Accessed on: 8 Jun. 2026.

QIAN, Yanan; ZHENG, Qinghua; SAKAI, Tetsuya; YE, Junting; LIU, Jun. **Dynamic author name disambiguation for growing digital libraries.** Information Retrieval Journal, v. 18, n. 5, p. 379–412, 2015.

RODRIGUES, Maria Eduarda Pereira; RODRIGUES, António Moitinho. **Desambiguação de nomes de autores: um desafio para os repositórios / Disambiguation of authors' names: a challenge for repositories.** Ciência da Informação, v. 53, n. 3, 2024.

RODRIGUES, Natan de Souza; COSTA, Aurelio Ribeiro; LEMOS, Lucas Correa; RALHA, Célia Ghedini. **Multi-strategic approach for author name disambiguation in bibliography repositories.** In: ANNUAL INTERNATIONAL CONFERENCE ON INFORMATION MANAGEMENT AND BIG DATA, 2020. Proceedings [...]. Cham: Springer, 2020. p. 63–76.

RODRIGUES, Natan de S.; RALHA, Célia Ghedini. **A flexible and configurable system to author name disambiguation.** IEEE Access, p. 1–1, 2025. Early Access.

RODRIGUES, Natan S.; MARIANO, Ari M.; RALHA, Celia G. **Author name disambiguation literature review with consolidated meta-analytic approach.** International Journal on Digital Libraries, p. 765–785, 2024.

RODRIGUES, Natan S.; RALHA, Célia G. **A novel framework with ComMAND: a combined method for author name disambiguation.** Information Processing & Management, v. 63, n. 1, p. 104304, 2026.

SUBRAMANIAN, S.; KING, D.; DOWNEY, D.; FELDMAN, S. **S2AND: a benchmark and evaluation system for author name disambiguation.** In: ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES, 2021. Proceedings [...]. Los Alamitos: IEEE Computer Society, 2021. p. 170–179.

TANG, Jie; FONG, Alvis C. M.; WANG, Bo; ZHANG, Jing. **A unified probabilistic framework for name disambiguation in digital library.** IEEE Transactions on Knowledge and Data Engineering, v. 24, n. 6, p. 975–987, 2012.

ZHANG, Li; LU, Wei; YANG, Jinqing. **LAGOS-AND: a large gold standard dataset for scholarly author name disambiguation.** Journal of the Association for Information Science and Technology, v. 74, n. 2, p. 168–185, 2023.

ZHANG, Yutao; ZHANG, Fanjin; YAO, Peiran; TANG, Jie. **Name disambiguation in AMiner: clustering, maintenance, and human in the loop.** In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING, 24., 2018, New York. Proceedings [...]. New York: ACM, 2018. p. 1002–1011.

ZHOU, Anji; SHI, Minghui; YUAN, Rui. **An effective author name disambiguation framework for large-scale publications.** IEEE Access, v. 12, p. 182086–182100, 2024.