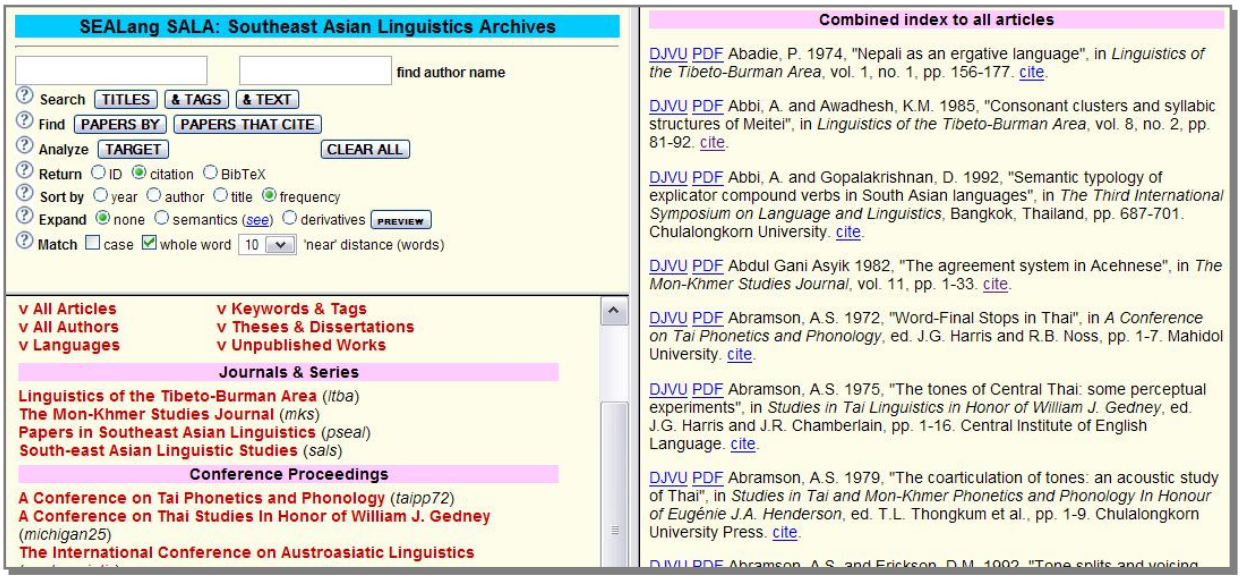
# SEAlang SALA: the Southeast Asian Linguistics Archives

*Doug Cooper, CRCL*
*http://sealang.net/sala*

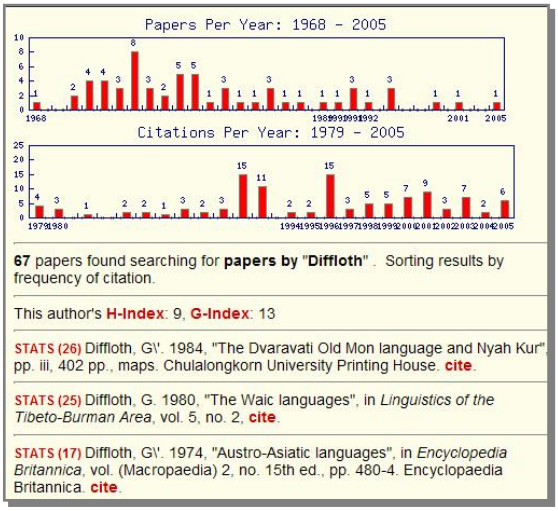
**ABOUT THE PROJECT** The *SEAlang SALA* collects, scans, indexes, and disseminates scholarly publication on Southeast Asian languages and linguistics, and devises and tests innovative approaches to aggregating and assisting discovery of the field's scattered literature. The SALA includes:



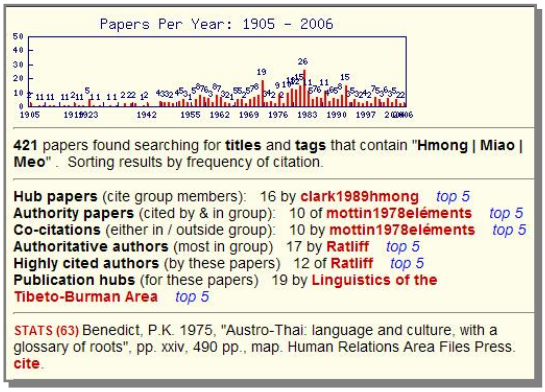

- A *searchable on-line archives* of the full content of most of the major journals, conference proceedings, series, and collections that focus on (mainland) Southeast Asian linguistics, as well as much unpublished material (field notes, theses) that is unavailable in the US.
- A *bibliography* (in *BibTex* and reusable XML-tagged formats) of related language and linguistics publication. This provides much deeper coverage than existing alternatives, extends well beyond mainland SEA, and plays an essential part in our research on automated citation disambiguation.
- A *corpus of citations* and *analytical citation index*, with Web-based tools that assist in discovery and evaluation of texts (*who does this article cite? who cites this article? what articles have similar citations?*), and calculates impact factor, G-index, and other measures of influence.
- An *API* that allows computer-to-computer interoperability between SALA resources and other online tools (e.g. *Multi-Tree*, *LL-MAP*, and our own *SEAlang Library* and *Mon-Khmer Languages* projects), as well as data sharing and citation tracking with other digital archives.

**AIDING DISCOVERY** Southeast Asia is one of the most linguistically diverse regions on the planet, with many, many hundreds of languages from five major families: Austroasiatic, Tai-Kadai, Hmong-Mien, Sino-Tibetan, and Austronesian. This surfeit of riches has an immediate consequence for the researcher and reference librarian: appropriate resources can be terribly difficult to locate, even with well-indexed reference data, simply because:

> "*the range of linguistic and cultural groupings within Southeast Asia is so great that even those who have studied the region for an academic lifetime can only acquire real competence in a limited area.*" Barbara Watson Andaya, *Historian*, Spring 1995




*Citation graph analysis* comes to the rescue; unraveling the tangled ties between authors and publications (and proving that citation is indeed the sincerest form of flattery)! Every reference in every article is individually indexed, using experimental software we wrote to help disambiguate the multitude of not-quite-identical entries that must ultimately point to the same canonical item. This preliminary project is just starting the lengthy data collection and correction process, but the SALA already demonstrates the technique's extraordinary power:




- **Hub papers** cite many of their fellows from the current group. They are often survey papers, and provide the best overall introduction to a field.
- **Authority papers** are sources within the current group that are cited most often by other papers in the group. They're usually the most important papers on this topic.
- **Co-citations** are cited by group members, but (unlike authority papers) aren't required to *be* members. They are often pointers to fundamental background reading.
- **Authoritative authors** have the most papers in the group – they write most frequently about this topic.
- **Highly cited authors** wrote the papers that are cross-referenced most often by this group.
- **Publication hubs** are journals, conference proceedings, and books: this group's publication venues.

No one approach meets all needs; every approach serves *some* need. When combined with full-text search and semantic query expansion (below), the SALA is a remarkably effective research assistant.

**MEASURING INFLUENCE** Journals have long used measures like *impact factor* to estimate influence. The SALA calculates metrics like the **H-index** (the number of papers, *h*, cited at least *h* times) and **G-index** (the number of papers, *g*, collectively cited $g^2$ times), which can be equally useful tools for authors *and their departments* in demonstrating the broad impact of SEA research.

**SEMANTIC QUERY EXPANSION** The SALA's facilities for extending queries are of special practical and research interest. Individual Southeast Asian language (and even families) are identified in the literature by a dizzying variety of names; sometimes due to simple spelling variation, but often the result of differences between speaker self-identification and (often derogatory) outsider names. However, we can predict many likely alternatives, and automatically build extended match groups of (contextually) semantic equivalents. The **Analyze target** search tool adds to this functionality, by subgrouping returned references according to the exact matched pattern. This is rather helpful in understanding when and why particular names and terminology have come and gone.

The expanded form of query "**Nyahkur**" is "**(Nyahkur|Nyah Kur|Nyakur|Niakuol|Niakuoll|Nyahkur Petchabun|Chaobon|Chaodon)**"

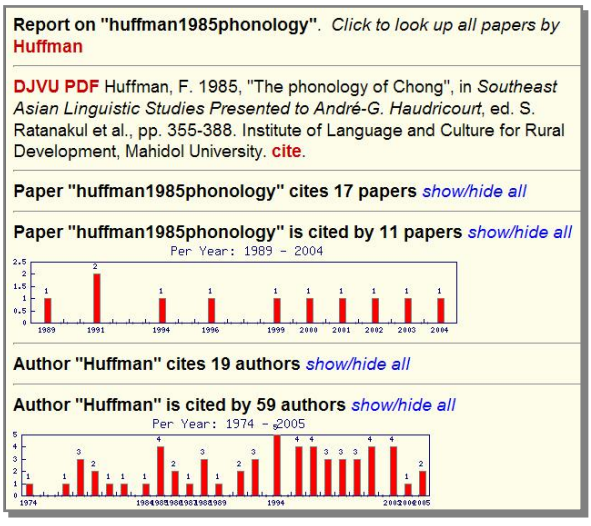
The expanded form of query "**writing thai**" is "**(orthography|orthographic|writing|script) (thai|siamese|siam)**"

**PRESERVATION AND ACCESS** Free, on-line access to publication is essential to the health of the field for many reasons: the unusual extent to which basic data appear in journals and conference proceedings, the increasing difficulty and cost of obtaining materials published outside of the US, and the equal difficulty faced by our colleagues in Southeast Asia who cannot access expensive aggregator services like JSTOR. The literature plays a critical role in educating young scholars – our future colleagues – in the region, and it is incumbent upon us to help stop (in Carol Mitchell's memorable phrase from the 65th *IFLA* conference), "serial murder in Southeast Asia."

The SEAlang SALA has had exceptional success in persuading intellectual property owners to allow open access to content. Besides scanning, indexing, and hosting the texts (and, increasingly, the journal sites as well, e.g. *MKS*, *SEALS*), we help publications make the most effective case they can for support by:

- giving them extensive, detailed statistics on when, where, and how their works are cited, and
- helping increase these figures by making it as easy as possible to discover *and cite* content.

These provide powerful arguments for institutional backing as well as subscriptions, and help all of us – authors, readers, and publishers – find a way to join forces in supporting ready access to printed knowledge.




*The SEAlang SALA is in preliminary development. Suggestions on design, content, and services are welcome; please contact* **doug.cooper.thailand@gmail.com**. *The SALA is committed to* ***open access*** *and* ***interoperability****, and is supported by the Center for Research in Computational Linguistics, a US 501(c)3 nonprofit.*