EFEO-CNRS-SOAS word list for linguistic fieldwork in Southeast Asia

Frederic Pain, Michel Ferlus, Alexis Michaud, Thị Thu Hà Phạm, Ryan Gehrmann, Minh-Châu Nguyễn

EFEO-CNRS-SOAS Word List
for Linguistic Fieldwork in Southeast Asia

Version 3

Lexique EFEO-CNRS-SOAS
pour les enquêtes linguistiques en Asie du Sud-Est

Version 3

法國遠東學院、法國國家科學研究院與倫敦亞非研究學院 東南亞語言調查詞彙表
第 3 版

Bảng từ EFEO-CNRS-SOAS
Dùng cho nghiên cứu điền dã ngôn ngữ học ở Đông Nam Á

bản 3

2019

EFEO-CNRS-SOAS Word List
for Linguistic Fieldwork in Southeast Asia

This word list aims to allow researchers (i) to conduct in-depth lexical investigation when doing fieldwork on languages of Southeast Asia, and (ii) to navigate between languages and dialects, through the use of a unique identifier for each lexical entry.

The earliest version of this word list was created by Ecole Française d'Extrême-Orient (EFEO, directed at that time by Georges Cœdès). The EFEO printed this list as a leaflet which was entrusted to civil servants of the colonial administration (*Questionnaire linguistique*, Hanoi: Imprimerie d'Extrême-Orient, 1938), aiming at extensive coverage of language varieties. The investigation was launched in 1938. Filled leaflets were gradually gathered at EFEO in Hanoi, until the process was interrupted by the war (in 1940). This collection constitutes one of the main bases of Haudricourt's comparative linguistic research during his stay at EFEO in Hanoi at the end of the 1940s.[1]

An enriched version of the word list was elaborated at the CNRS laboratory CeDRASEMI (Centre de documentation et de recherche sur l'Asie du Sud-Est et le monde insulindien). This version has an added digit for additional words: for instance, in addition to "11. lighting" were added "11.1 it lightens" (i.e. "there are flashes of lightning") and "11.2 the lightning strikes". This overhaul was supervised by Lucien Bernot, probably between 1960 et 1970. The linguistic questionnaire is described as "jointly prepared by the Centre de documentation et de recherche sur l'Asie du Sud-Est et le monde insulindien (EPHE-CNRS, Paris) and the Department of South Asia and Oceania of the School of Oriental Studies (University of London) with a view to creating an Ethnolinguistic Atlas of Southeast Asia".

Michel Ferlus re-typed the 22-page list to adopt a format suitable for use in the field. As the list remained insufficiently comprehensive for in-depth linguistic fieldwork, Michel Ferlus added further items in the course of his field trips to Vietnam in the 1990s, and asked Nguyễn Phú Phong to provide Vietnamese glosses. The list was later reworked by Trần Trí Dõi, who replaced Southern Vietnamese vocabulary by Northern Vietnamese words, better adapted to the fieldwork locations that Michel Ferlus was exploring in collaboration with him.

This list was circulated among Michel Ferlus's colleagues and collaborators. Khmer glosses were added by Frédéric Pain, based on a version of the CeDRASEMI-SOAS list to which Marie Martin had added Khmer glosses.

The list is maintained by Alexis Michaud.

---

[1] This investigation is mentioned on page 308 of Haudricourt's 1956 article "De la restitution des initiales dans les langues monosyllabiques : le problème du thai commun", *Bulletin de la Société de Linguistique de Paris* 52. 307–322.

Version history:

About version 1 (2014):

The main adjustments made in 2013-2014 (at the International Research Institute MICA, HUST - CNRS/UMI-2954 - Grenoble INP) were the following:

- converting from word-processor file (MS-Word) to spreadsheet format
- adding a new numbering (indicated as 'UID', for Unique IDentifier) to facilitate annotation of corresponding sound files[2]
- adding a label which (redundantly) indicates the layer to which the entry belongs: the entries that date back to the 1938 EFEO version carry the label "I"; those added in the CNRS-SOAS version carry the label "II"; those added by Michel Ferlus carry the label "III"; and those added by Frédéric Pain carry the label "IV". (The label "V" is used for custom items added by users of the word list when they study a new language variety.)
- adding Chinese glosses, supplementing the English glosses, and revising the Vietnamese glosses (work conducted by Phạm Thị Thu Hà and Alexis Michaud)

About version 2 (2016):

Ryan Gehrmann supervised improvements to the English glosses. The French > English translations of the glosses were divided up and sent out to friends and acquaintances of RG who are native speakers of English and have experience in French. They were encouraged to use dictionaries and other online resources to double check the accuracy of the translations and also, wherever necessary, to improve the naturalness of the translations. The following individuals contributed to this effort:

- Hitoshi Yamaguchi (200 items)
- Linnea Tideman (200 items)
- Lynne O'Connor (200 items)
- Robert Lasher (200 items)
- John Berthelette & Doug Frasier (400 items)
- Ryan Gehrmann (1440 items)

Nguyễn Thị Minh Châu made some improvements to the Vietnamese glosses.

About version 3 (2019):

In Version 3, Ryan Gehrmann and his collaborators have added glosses for four major languages of the region: Burmese, Central Thai, Northern Thai and Lao. The translators worked off of the English translation of the original French glosses. The new glosses in these four languages remain in draft form and they have not yet been field tested as of early 2019. Additionally, the Lao translation is not yet complete. The editors of the word list would value and appreciate any feedback, corrections or additions for the new glosses. If you do have any such feedback, please send it to Ryan Gehrmann (ryangehrmann@gmail.com).

[2] Note that UID 1488 and UID 1644 are empty lines: the items at issue were doublets, which were removed. The lines are retained for technical convenience, to allow easy copy/paste from and to the spreadsheet.

We must thank the following individuals for the hard work that they put in to developing the new gloss translations.

- Mr. Lin Kyaw Zaw (Burmese)
- Ms. Sutthinee Promkandorn (Central & Northern Thai)
- Ms. Leah Doty (Lao)

Future work:

This list is a working tool. Among other possibilities for future improvements:

- The Chinese translations would benefit greatly from another round of verification by a native speaker who also understands the original (French) glosses.
- A systematic syntax could be adopted: items such as 'to spread out the mat', which include both a verb and an object noun, could be rewritten systematically as 'to spread out (the mat)', parenthesize the incidental part of the elicitation, and indicating, in a usage note for the list, that the form in the target language should also use parentheses if possible, to allow for convenient extraction of the noun and verb by end users. Indications on part of speech would be another useful addition.
- Pictures and videos would make it much easier to elicit vocabulary related to the flora and fauna, as well as for other parts of the lexicon.

The word list is offered online in Open Office format (.ods) and MS-Excel format. The entire structure of the spreadsheet is as follows.

column	**contents**
A	UID: Unique IDentifier assigned in 2014
B	Michel Ferlus's numbering
C	layer to which the entry belongs: I = EFEO; II = SOAS +CNRS; III = Michel Ferlus; IV = Frédéric Pain
D	French gloss
E	English gloss
F	Mandarin Chinese gloss, in simplified characters
G	Burmese gloss
H	Central Thai gloss
I	Northern Thai gloss
J	Lao gloss
K	Vietnamese gloss
L	Written Khmer gloss
M	Standard Modern Khmer in IPA
N	Tatey Leu dialect of Khmer
O	indication of semantic field, in English
P	indication of semantic field, in French
Q	Target language

Many thanks to the colleagues who contributed corrections, suggestions and encouragement, in particular Doug Cooper, Cao Thành Việt, Nathan Hill, Minh Châu Nguyễn and Justin Watkins.

This word list can be cited as:

> Frédéric Pain, Michel Ferlus, Alexis Michaud, Phạm Thị Thu Hà, Ryan Gehrmann (2019). EFEO-CNRS-SOAS word list for linguistic fieldwork in Southeast Asia, version 3. Identifier: oai:halshs.archives-ouvertes.fr:halshs-01068533. Available: http://halshs.archivesouvertes.fr/halshs-01068533.

Lexique EFEO-CNRS-SOAS
pour les enquêtes linguistiques en Asie du Sud-Est

Cette liste numérotée constitue un outil pour l'élicitation de vocabulaire. Elle vise également à permettre aux chercheurs de naviguer entre les langues et les dialectes recueillis au fil des ans et sur tous les terrains d'Asie du Sud-Est.

La première version de ce lexique a été élaborée par l'Ecole Française d'Extrême-Orient (EFEO), alors dirigée par Georges Cœdès, pour une vaste enquête lancée en 1938 et interrompue par la guerre en 1940. L'EFEO en a imprimé une quantité sous la forme de petits fascicules, distribués aux fonctionnaires envoyés en mission par l'administration coloniale (*Questionnaire linguistique*, Hanoi: Imprimerie d'Extrême-Orient, 1938). Ces fascicules, remplis avec plus ou moins de talent, constituent l'une des principales bases des travaux comparatistes réalisés par Haudricourt lors de son séjour à l'EFEO de Hanoi, à la fin des années 1940 (voir Haudricourt 1956:308).

Une version enrichie a été élaborée au laboratoire CeDRASEMI du CNRS (Centre de documentation et de recherche sur l'Asie du Sud-Est et le monde insulindien). Lucien Bernot a été la cheville ouvrière de cette amélioration qui a dû se faire entre 1960 et 1970. Cette version était décrite comme un "Questionnaire linguistique préparé conjointement par le Centre de documentation et de recherche sur l'Asie du Sud-Est et le monde insulindien (EPHE-CNRS, Paris) et le Département d'Asie du Sud-Est et d'Océanie de la School of Oriental and African Studies (University of London) en vue de l'établissement d'un Atlas ethnolinguistique de l'Asie du Sud-Est". Par rapport au livret de 1938, la version CNRS-SOAS est augmentée des nombres avec décimales: 11.1 11.2 etc.

Michel Ferlus a re-tapé cette liste pour en faire des cahiers d'enquête commodes à remplir sur le terrain. Comme cette liste restait insuffisante pour une bonne utilisation linguistique, Michel Ferlus, au cours de ses enquêtes au Vietnam dans les années 90, a augmentée cette liste des nombres en *italique*, et a fait établir une traduction en vietnamien par Nguyễn Phú Phong ; cette liste a ensuite été remaniée par Trần Trí Dõi, qui a remplacé les mots de vocabulaire du dialecte du sud du Vietnam par le vocabulaire du nord du pays, plus adapté aux terrains que Michel Ferlus explorait avec lui à cette époque.

Cette liste a circulé parmi les collègues et collaborateurs de Michel Ferlus. Elle a été enrichie de gloses en khmer par Frédéric Pain, en partie fondées sur la version de la liste CeDRASEMI-SOAS annotée en khmer par Marie Martin.[3]

[3] Les traductions khmères consistent en : (1) une translitération du khmer écrit ; (2) une notation (phonétique) du khmer central, dit standard, tirée du dictionnaire d'Antelme et Bru-Nut (2001) ; (3) une notation (phonétique) du dialecte khmer des Cardamomes, récolté par Marie Alexandrine Martin (Martin 1969; voir un article au sujet du parler de Tatey: Martin 1975). Parmi les données présentes dans la liste de M.A. Martin, le dialecte parlé dans la région de Tatey Lœ a été choisi (plutôt que les dialectes parlés à Russey Chrum, Chke Prus et Thung Krang) car c'est celui qui atteste le plus d'archaïsmes prenant leurs racines en moyen khmer. Le khmer de Tatey Lœ maintient l'opposition des phonèmes [ɨ]-[i]-[iː] du moyen-khmer par une opposition de trois paires registrales distinctes alors que le khmer du nord n'atteste plus que deux oppositions et le khmer central une

Le présent document est maintenu par Alexis Michaud (alexis.michaud@cnrs.fr).

Historique des versions:

- **Version 1 (2014):**

Les ajustements principaux effectués à l'Institut de recherche international MICA (HUST - CNRS/UMI-2954 - Grenoble INP) en 2013-2014 ont été les suivants:

- le document a été converti au format tableur (Open Office et MS Excel)
- une nouvelle numérotation a été établie, pour faciliter l'emploi de la liste
- des gloses en chinois ont été ajoutées ; les gloses en anglais ont été complétées ; et les gloses en vietnamien ont été intégralement révisées par Phạm Thị Thu Hà et Alexis Michaud.

- **Version 2 (2016):**

Ryan Gehrmann a coordonné un travail d'équipe qui a permis de revoir intégralement les gloses anglaises.

Nguyễn Thị Minh Châu a apporté des améliorations aux gloses vietnamiennes.

- **Version 3 (2019):**

Ryan Gehrmann et ses collaborateurs ont ajouté des gloses pour quatre langues majeures de la région : birman, thaï central (siamois), thaï du nord et laotien. Les traducteurs ont travaillé à partir de la traduction anglaise des gloses originales françaises. À la date du 18 mars 2019, les gloses dans ces quatre langues sont encore à l'état d'ébauche et n'ont pas encore été testées sur le terrain. En outre, la traduction laotienne n'est pas encore terminée. Les éditeurs de la liste de mots apprécieraient tout commentaire, correction ou ajout pour améliorer ces gloses. Les informations sont à envoyer à Ryan Gehrmann (ryangehrmann@gmail.com).

Des remerciements tout particuliers sont dûs aux personnes suivantes pour le travail acharné qu'elles ont consacré à l'élaboration des nouvelles traductions :

- M. Lin Kyaw Zaw (birman)

- Mme Sutthinee Promkandorn (thaï central et thaï du nord)

- Mme Leah Doty (lao)

seule. Ainsi, en khmer des Cardamomes, les trois phonèmes moyen-khmer [ɨ]-[i]-[iː] ne se sont pas confondus et représentent bien chacun une paire registrale indépendante; en khmer du nord, il y a eu confusion de deux voyelles moyen-khmer [ɨ]-[i] en une seule paire registrale [ʌ]-[ɯ] en opposition avec la paire registrale [e]-[i] (provenant du moyen-khmer [iː]); en khmer central, les trois phonèmes moyen-khmer se sont confondus en une seule paire registrale [ɤ]-[ɨ]. Le second phonème de la paire registrale est celui du second registre.

moyen khmer	*khmer des Cardamomes* (Tatey Lœ)	*khmer du nord* (Surin)	*khmer central* (standard)
[ɨ]	[ə]-[ɨ]	[ʌ]-[ɯ]	[ɤ]-[ɨ]
[i]	[e]-[i]	[ʌ]-[ɯ]	[ɤ]-[ɨ]
[iː]	[e]-[i]	[e]-[i]	[ɤ]-[ɨ]

Pour plus de détails, voir Ferlus (1992).

Bảng từ EFEO-CNRS-SOAS

Dùng cho nghiên cứu điền dã ngôn ngữ học ở Đông Nam Á

Đây là một bảng từ được lập ra nhằm mục đích giúp người nghiên cứu đối chiếu giữa các ngôn ngữ và các phương ngữ được thu thập trong quá trình làm việc điền dã ở Đông Nam Á.

Phiên bản đầu tiên của bảng từ này do Ecole Française d'Extrême-Orient (Viện Viễn đông Bắc cổ, tên viết tắt là EFEO) lập ra (dưới sự điều hành của Georges Cœdès) để phục vụ cho một cuộc điều tra trên diện rộng nghiên cứu về sự dạng ngôn ngữ được bắt đầu năm 1938 và bị ngắt quãng năm 1940 vì lí do chiến tranh. EFEO đã in bảng từ này dưới dạng bảng hỏi về ngôn ngữ và phân phát cho các công chức của chính quyền thực dân (*Questionnaire linguistique*, Hanoi: Imprimerie d'Extrême-Orient, 1938). Những tờ bảng hỏi thoạt đầu còn thô sơ này đã làm nên một trong những nền tảng cơ bản trong nghiên cứu ngôn ngữ học đối chiếu của Haudricourt trong thời gian ông ở Hà Nội vào cuối những năm 40 (tham khảo Haudricourt 1956:308).

Phiên bản cải tiến của bảng từ này đã được phát triển tại phòng thí nghiệm CeDRASEMI (Centre de documentation et de recherche sur l'Asie du Sud-Est et le monde insulindien) thuộc Trung tâm Nghiên cứu Khoa học Quốc gia Pháp (CNRS). Phiên bản này sử dụng số thập phân để đánh dấu những từ ngữ được bổ sung thêm về sau. Ví dụ: với tư "11. lighting" đã có (trong phiên bản gốc), người ta thêm vào "11.1 it lightens" và "11.2 the lightning strikes". "11.1" và "11.2" cho biết đây là hai từ ngữ đã được thêm vào trong phiên bản cải tiến. Công cuộc đại tu này được thực hiện dưới sự giám sát của Lucien Bernot trong khoảng thời gian giữa những năm 1960 và 1970. Bảng hỏi ngôn ngữ học (linguistic questionnaire) được miêu tả như một "sự chuẩn bị của Trung tâm Tư liệu và Nghiên cứu Đông Nam Á và khu vực Nam đảo (EPHE-CNRS, Paris) hợp tác với Khoa Nam Á và Châu Đại Dương – Trường Đông phương học – Đại học Luân Đôn, nhằm hướng đến việc lập ra tập atlas về ngôn ngữ học nhân học (Ethnolinguistic) ở Đông Nam Á".

Michel Ferlus đã đánh lại bảng từ gồm 22 trang này với một hình thức phù hợp hơn cho nghiên cứu điền dã. Nhận thấy bảng từ chưa thực sự toàn diện để sử dụng cho việc đi sâu nghiên cứu ngôn ngữ, trong quá trình nghiên cứu điền dã ở Việt Nam những năm 90, Michel Ferlus đã tiếp tục bổ sung các mục từ rộng hơn và đã đề nghị Nguyễn Phú Phong bổ sung phần dịch nghĩa tiếng Việt. Bảng từ này về sau được Trần Trí Dõi biên tập lại. Các từ ngữ Nam bộ được thay thế bằng các từ ngữ Bắc bộ, để tiện hơn cho công tác điền dã ở khu vực mà Michel Ferlus mở rộng nghiên cứu với sự cộng tác của Trần Trí Dõi.

Bảng từ này được lưu hành giữa các đồng nghiệp và cộng tác viên của Michel Ferlus. Frédéric Pain đã thêm vào phần dịch nghĩa tiếng Khmer, dựa trên bảng từ CeDRASEMI-SOAS vốn đã được Marie Martin bổ sung phần nghĩa tiếng Khmer trước đó. Bảng từ tiếp tục được cập nhật tại viện nghiên cứu quốc tế MICA (HUST - CNRS/UMI-2954 - Grenoble INP) trong hai năm 2013-2014. Hiện nay, phiên bản mới nhất đang được Alexis Michaud lưu trữ.

Quá trình cải tiến các phiên bản:

Phiên bản 1 (năm 2014)

Những điều chỉnh chính được thực hiện trong năm 2013-2014 bao gồm:

- Đổi từ định dạng văn bản (MS-Word) sang định dạng bảng biểu (spreadsheet).
- Thêm cột số thứ tự định danh được đánh liên tiếp nhau để thuận tiện hơn cho việc chú thích các tập tin âm thanh. Cụ thể, mỗi một mục từ sẽ được gán một số định danh riêng biệt được gọi là 'UID', viết tắt của *Unique Identifier*.
- Thêm cột nhãn bằng số la mã cho biết thứ tự xuất hiện của các mục từ trong bảng từ: các mục từ có từ phiên bản đầu tiên của EFEO năm 1938 mang nhãn "I"; các mục từ được bổ sung trong phiên bản cải tiến của CNRS-SOAS mang nhãn "II"; các mục từ được thêm bởi Michel Ferlus mang nhãn "III"; và phần bổ sung của Frédéric Pain mang nhãn IV. (Nhãn "V" dành cho trường hợp các mục từ có thể tiếp tục được thêm vào sau này bởi người nghiên cứu trong quá trình họ sử dụng bảng từ để tìm hiểu về một ngôn ngữ mới)
- Thêm phần dịch nghĩa tiếng Trung Quốc; bổ sung những chỗ thiếu trong phần dịch nghĩa tiếng Anh và tiếng Việt; kiểm tra lại phần dịch nghĩa tiếng Việt đã có trong phiên bản cũ (công việc này do Phạm Thị Thu Hà và Alexis Michaud thực hiện).

Phiên bản 2 (năm 2016):

Ryan Gehrmann là người đã giám sát các cải tiến cho tiếng Anh. Phần dịch từ tiếng Pháp sang tiếng Anh đã được chia nhỏ ra và gửi cho bạn bè và người quen của Ryan, những người nói tiếng Anh bản ngữ và có kinh nghiệm về tiếng Pháp. Họ được khuyến khích sử dụng từ điển và các tài nguyên trực tuyến khác để kiểm tra tính chính xác của các bản dịch và chỉ ra những chỗ cần cải thiện tính tự nhiên. Những người dưới đây đã đóng góp trong việc này:

• Hitoshi Yamaguchi (200 từ)

• Linnea Tideman (200 từ)

• Lynne O'Connor (200 từ)

• Robert Lasher (200 từ)

• John Berthelette & Doug Frasier (400 từ)

• Ryan Gehrmann (1440 từ)

Nguyễn Thị Minh Châu đã thực hiện một số cải tiến cho tiếng Việt.

Phiên bản 3 (năm 2019):

Trong phiên bản 3, Ryan Gehrmann và các cộng tác viên của mình đã thêm bản dịch cho bốn ngôn ngữ chính của khu vực: Miến Điện, Trung Thái, Bắc Thái và Lào. Các dịch giả đã làm việc trên bản dịch tiếng Anh của bản gốc tiếng Pháp. Bảng từ tương ứng cho bốn ngôn ngữ này hiện còn đang ở dạng bản nháp và chúng vẫn chưa được biên tập lại tính tới đầu năm 2019. Ngoài ra, bản dịch tiếng Lào vẫn chưa hoàn thành. Ban tập viên hiện tiếp nhận và đánh giá cao mọi phản hồi, đề xuất chỉnh sửa cho các bản dịch tương ứng của bốn ngôn ngữ mới này. Nếu bạn có bất kỳ ý kiến đóng góp nào, xin vui lòng liên hệ Ryan Gehrmann (ryangehrmann@gmail.com).

Chúng tôi chân thành cảm ơn các cá nhân sau đây vì đã làm việc chăm chỉ để góp phần phát triển các bảng từ tương ứng trong các ngôn ngữ hiện đang được bổ sung.

• Thầy Lin Kyaw Zaw (Miến Điện)

• Cô Sutthinee Promkandorn (Trung & Bắc Thái)

• Cô Leah Doty (Lào)

References / Références citées / Tài liệu tham khảo:

Antelme, Michel Rethy & Hélène Suppya Bru-Nut. 2001. *Dictionnaire français-khmer*. Paris: L'Asiathèque.

Ferlus, Michel. 1992. Essai de phonétique historique du khmer (du milieu du premier millénaire de notre ère à l'époque actuelle). *Mon-Khmer Studies* 21. 57–89.

Haudricourt, André-Georges. 1956. De la restitution des initiales dans les langues monosyllabiques : le problème du thai commun. *Bulletin de la Société de Linguistique de Paris* 52. 307–322.

Martin, Marie A. 1969. Vocabulaire du khmer des Cardamomes: Tatey Lœ, Russey Chrum, Chke Prus et Thung Kran.

Martin, Marie A. 1975. Le dialecte cambodgien parlé à Tatey, massif des Cardamomes. *Asie du Sud-Est et Monde Insulindien* 6(4). 71–79.