Interpersonal meaning annotation for Asian language corpora: The case of TUFS Asian Language Parallel Corpus (TALPCo)

野元　裕樹　　岡野　賢二

スニサー・ウィッタヤーパンヤーノン　　野村　純太

東京外国語大学

言語処理学会第25回年次大会、2019年3月14日

@名古屋大学

東京外国語大学
Tokyo University of Foreign Studies

本発表の目的

- アジア言語では、**自然な会話文の実現には対人的意味の情報が重要**であることを示す。
- 日本語との翻訳システムの開発に有益と思われる、**東南アジア言語**の特徴を紹介する。
- 東京外大アジア言語パラレルコーパス（TUFS Asian Language Parallel Corpus; TALPCo）への対人的意味情報のアノテーションについて報告する。
  - より大規模な言語資源への足掛かりに
  - 会話の自然さ、翻訳の精度の評価尺度への組み入れ
  - 通言語的に使える対人的意味の分析法

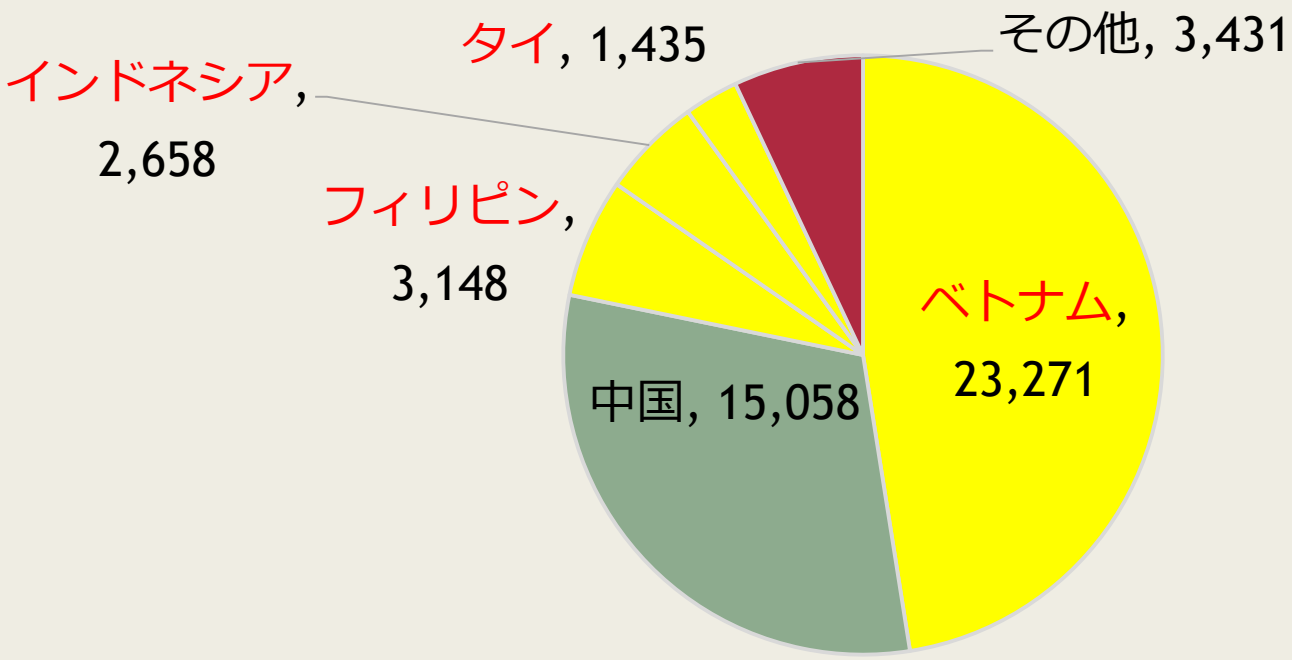
TALPCo (Nomoto et al. 2018)

- https://github.com/matbahasa/TALPCo
- NICTの Asian Language Treebank (ALT) Parallel Corpus (Riza et al. 2016) がお手本
- 収録言語
  - 日本語、ビルマ語、マレー語、インドネシア語、英語（2018年）
  - タイ語、ベトナム語（2019年） NEW
- 提供データ
  - 平文
  - トークン化された文
  - 複単語・複音節表現リスト
  - アラインメント NEW
  - 対人的意味情報 NEW

なぜ日本語×東南アジア言語か？

- 言語の研究・言語教育
- 社会的意義
  - 日系企業の東南アジア進出
  - 東南アジアからの訪日者の増加

JITCO入国支援**技能実習生**（１号）（2017年10月末分）



国際交流基金「2015年度**日本語教育**機関調査結果概要」学習者数順位

1	中国	953,283
2	インドネシア	745,125
3	韓国	556,237
4	オーストラリア	375,348
5	台湾	220,045
6	タイ	173,817
7	アメリカ	170,998
8	ベトナム	64,863
9	フィリピン	50,038
10	マレーシア	33,224

JASSO2017年度外国人**留学生**在籍状況調査結果

1	中国	107,260
2	ベトナム	61,671
3	ネパール	21,500
4	韓国	15,740
5	台湾	8,947
6	スリランカ	6,607
7	インドネシア	5,495
8	ミャンマー	4,816
9	タイ	3,985
10	マレーシア	2,945

TALPCo データの特徴

- TUFS言語モジュールの語彙モジュールに利用されているデータ
- 日本語能力試験N5レベル（最も易しいレベル）
- 日常生活で使うような、短く、平易な文
- 1,372文
- フォーマルな話し言葉の文体

帰る　電車が　なかったので、　友達の　家に　泊まりました。
[3156]

- 翻訳も同じ文体

日本語データ

1176 田中さんは 学生では ありません。

1178 父は 先生です。

1180 学校は 休みです。

1194 東京は 晴れでした。

1222 公園に 木が あります。

1229 田中さんは どこに いますか。

1233 お金が ありません。

1244 机の 上に 本が あります。

1245 机の 下に かばんが あります。

1246 かばんの 中に ノートが あります。

パラレルな文の例

日： 学校は 休みです。

緬： ကျောင်းပိတ်တယ်။

馬： Sekolah cuti.

尼： Sekolah sedang libur.

泰： โรงเรียนหยุดครับ NEW

越： Trường nghỉ dạy. NEW

英： There is no school.

[1180]

タイ語、ベトナム語の追加によって露呈した問題

- 翻訳に必要な、文脈に関する情報が日本語文に不足。
  - 「わたし」／話し手は、男性？女性？
  - 「田中さん」は、男性？女性？
    話し手より年上？年下？どのくらい？
- 「～さん」の問題は、昨年、すでに認識。「マレー語に合わせる」という基本方針があったが…
  - それでも、言語間での不統一が生じた。
  - ベトナム語のように、マレー語よりも細かな区別をする言語に対応できない。

⇒翻訳の際、データの使用の際に、日本語文からは分からない文脈に関する情報を明示的にデータ化しよう！

タイ語の「です・ます」「わたし」

- 「です・ます」は、話し手の性により異なる。
  - 男性： ครับ khráp （ビルマ語も）
  - 女性： ค่ะ khâ （平叙文）、คะ khá（疑問文）
- 男女両方が使える「わたし」（フォーマルな会話の1人称代名詞）がない。

เมื่อวาน	ผม	เรียน	หนังสือ	ครับ
mɯ̂awaan	phǒm	rian	náŋsɯ̌ɯ	khráp
yesterday	I	learn	book	小辞

←男性が用いる表現

「きのう わたしは 勉強しました。」 [1356]

タイ語文を適切に使うための情報：話し手＝男性

ベトナム語の「～さん」

- 指示対象の性と話し手に対する相対的年齢差により細かく分けられる。

年齢層	弟・妹世代	兄・姉世代	親世代	祖父母世代
男性	em	anh	bác	ông
女性		chị		bà

Chị Kimura là học sinh nhưng anh Tanaka là nhân viên công ty.
Ms. Kimura is student but Mr. Tanaka is staff company
「木村さんは 学生ですが、 田中さんは 会社員です。」[3110]

ベトナム語文を適切に使うための情報：

- 木村さん＝女性、話し手の姉世代
- 田中さん＝男性、話し手の兄世代

対人的意味（interpersonal meaning)

- 話し手と聞き手、話し手と指示対象の間の関係性
- 文が適切に用いられるための条件を表す（使用条件的意味）
  cf. 真理条件的意味（＝「言いたいこと」）

[3289]	その コップは あなたのです。	That cup is yours.
客→店員：	適切	適切
店員→客：	正しいが、不適切	適切

- 「あなた」とyouは、真理条件は同じだが、使用条件が異なる。
- 使用条件を考慮しない→不適切（不自然、無礼）な文
  →人間関係☹

真理条件×

外界の事態を正しく記述できていない文

真理条件○ & 使用条件×

外界の事態を正しく記述するが、その発話状況での使用は不適切な文

真理条件○ & 使用条件○

外界の事態を正しく記述し、かつ、その発話状況での使用も適切な文

VoiceTra (Matsuda et al. 2013)
「そのコップはあなたのです。」

緬	ဒီခွက်ကမင်းဟာပါ။	မင်း mín対等または目下の聞き手「君」；インフォーマル
馬	Awak ada gelas itu.	「君はそのグラスがある」の意。 Awak: 対等または目下の聞き手「君」；インフォーマル
尼	Cankir itu milik Anda.	OK
泰	แก้วนั้นเป็นของคุณ	男性の声だが、ครับ khrápなし。書き言葉。 คุณ: 対等または目下の聞き手
越	Cái cốc của bạn là của bạn.	「あなたのコップはあなたのです」の意。 Bạn: 「友達」から転じた親しみを伴う「あなた」

Google翻訳
「そのコップはあなたのです。」

緬	အဆိုပါခွက်ကိုသင်တို့အဘို့ဖြစ်၏။	非文。「上述のコップをあなた方のためである。」 သင်တို့ tɪ̀ɴ-dò 文語体。発話されない形式。上下関係には関わらない。
馬	Cawan itu adalah milik anda.	anda: 普通、目の前にいない不特定多数の聞き手。
尼	Cawan itu milik Anda.	OK
泰	นั่นคือถ้วยของคุณ	女性の声だが、ค่ะ khâなし。書き言葉。 คุณ: 対等または目下の聞き手
越	Chiếc cốc đó là của bạn.	Bạn: 「友達」から転じた親しみを伴う「あなた」

Bing翻訳
「そのコップはあなたのです。」

緬	–	–
馬	Cawan adalah milik anda.	anda: 普通、目の前にいない不特定多数の聞き手。 「コップはあなたのです。」の意。
尼	Cangkir adalah milikmu.	-mu: 対等または目下の聞き手
泰	ถ้วยเป็นของคุณ	คุณ: 対等または目下の聞き手 女性の声だが、ค่ะ khâなし。 書き言葉。
越	Cốc là của cô.	cô: 話し手（／聞き手）の姉世代の女性 「コップはあなた（／私）のです。」の意。

TALPCoへの対人的意味情報付与

1. 語彙に対するアノテーション　data_言語名-IPLex.csv
2. 文脈に対するアノテーション
   a. 話し手　　data_言語名-IPSpkr.csv
   b. 聞き手　　data_言語名-IPAddr.csv

例：日本語－ベトナム語

teacher	先生		Thầy male
	、		ơi
	こちら		,
	が		đây
neutral, sg	私		là
	の		mẹ
female, parent, formal	母		của
	です		em student, sg
	。		.
neutral, sg		話し手	student, sg
teacher		聞き手	male, teacher

男性の「先生」
cf. 女性の「先生」cô

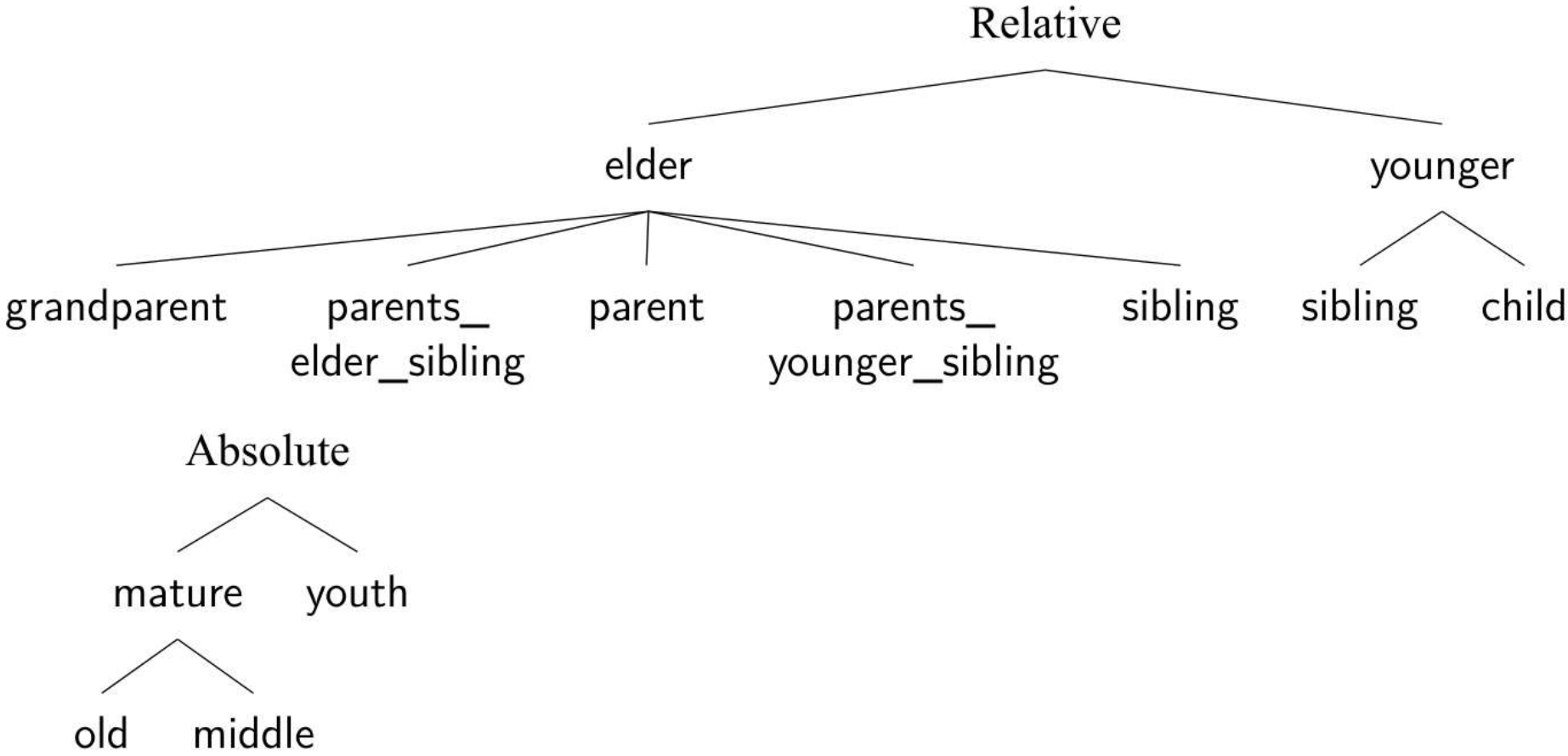
日本語の「母」の情報で十分

生徒としての「私」
cf. 中立的な「私」tôi

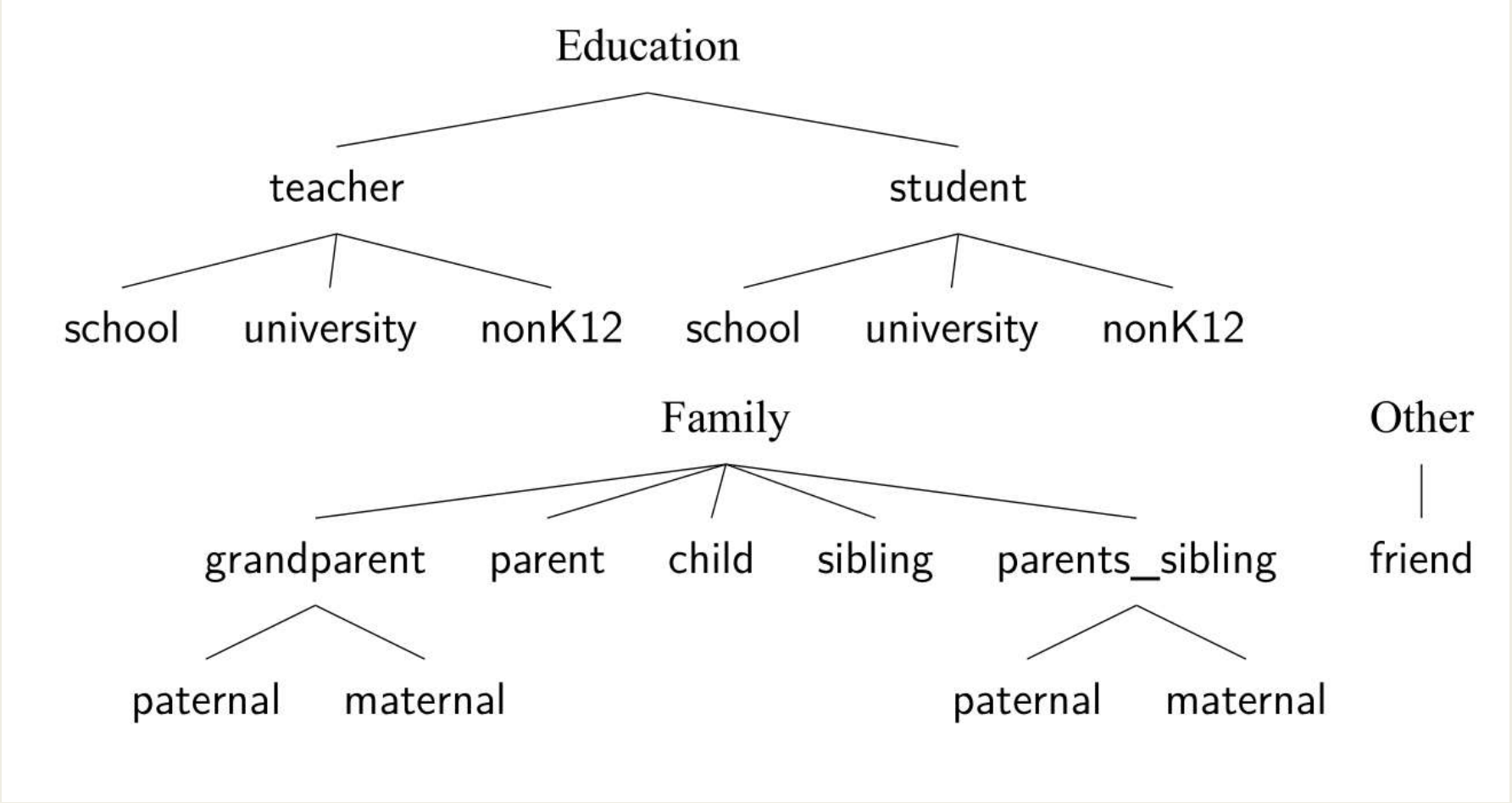
対人的意味の素性構造(1)：全体図（TALPCoで使用されている物のみ）

Gender（性）	male, female
Marital status（婚姻状況）	married, unmarried
Honour（敬意）	hon
Age（年齢）— Relative（相対的）, Absolute（絶対的）	
Social status（社会的地位）	higher, equal_or_lower, neutral
Role（役割）— Education（教育）, Family（家族）, Other（その他）	
Group（集団）	foreigner
Formality（格式度）	formal, informal
Number（数）	sg, pl, pl.incl, pl.excl

対人的意味の素性構造(2)：年齢



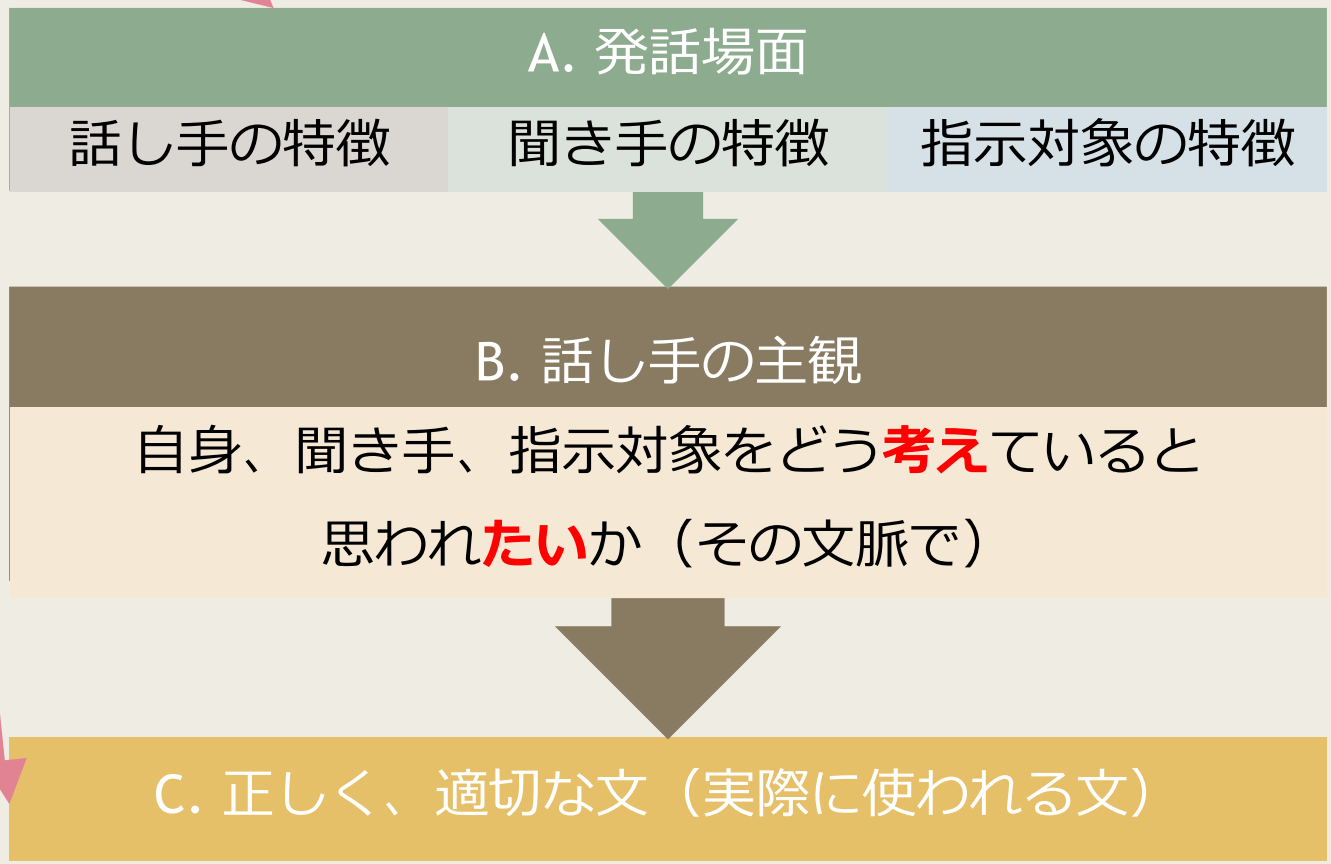
対人的意味の素性構造(3)：役割



実際の発話状況

客観的に観察可能

アノテーション付きコーパスから得られる情報



清水(2011: 135)「目下の人への気遣い」

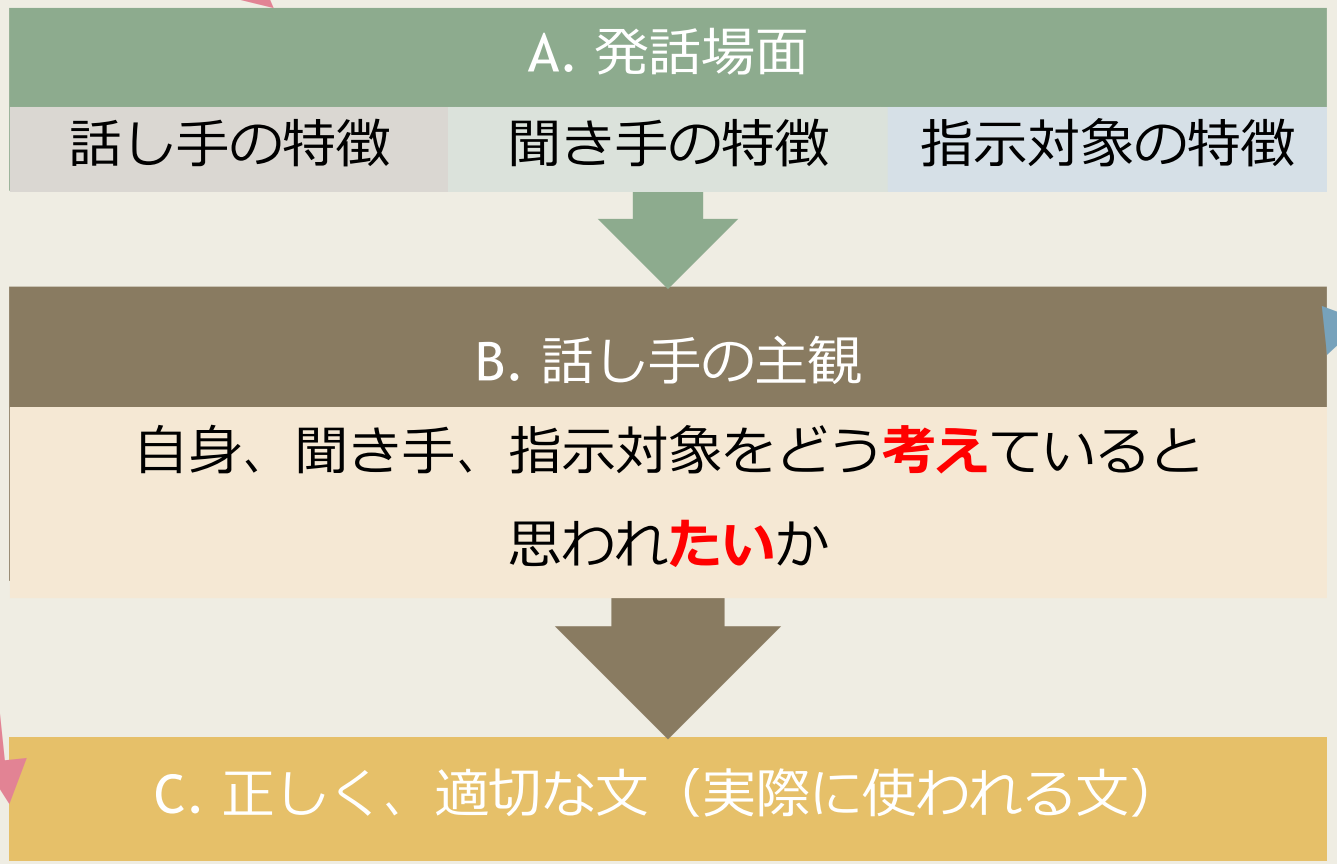
ベトナムで明らかに自分より年上の人と話をしているとき、しばしば自分が「叔父さん（chú）・叔母さん（cô）」、あるいは「伯父さん・伯母さん（bác）」と呼ばれることがある。一見戸惑いを感じるこの呼称法も実はベトナム人の細かい気遣いの表れで、明らかにその人達よりも年下である自分に対し、年下としての呼称「弟・妹（em）」や「甥・姪、孫（cháu）」等を直接用いること、つまり上からの目線で呼称することを避け、（中略）呼称した結果なのである。

実際の発話状況

客観的に観察可能

アノテーション付きコーパスから得られる情報

Bなしのモデルと A & Cから推論



今後の課題

- より多様な対人的意味を含むように、コーパスを拡張
- 対人的意味を表す表現の通言語的研究

  →素性構造の精緻化

  →汎用性
- フィリピノ語など他の東南アジア言語の追加

紹介
TUFS Open Language Resources

- TUFS言語モジュールの語彙モジュールの24言語のデータ
  - 語彙分類表*ID*
  - 単語
  - 例文
  - PostgresSQL database backup file (pg_dump file)
  - CC-BY 4.0
  - *https://malindo.aa-ken.jp/TUFSOpenLgResources.html*

東京外国語大学
Tokyo University of Foreign Studies

参考文献

- Matsuda, Shigeki, Xinhui Hu, Yoshinori Shiga, Hideki Kashioka, Chiori Hori, Keiji Yasuda, Hideo Okuma, Masao Uchiyama, Eiichiro Sumita, Hisashi Kawai, and Satoshi Nakamura. 2013. Multilingual speech-to-speech translation system: *VoiceTra*. *2013 IEEE 14th International Conference on Mobile Data Management* volume 2, 229-233.
- Nomoto, Hiroki, Kenji Okano, David Moeljadi and Hideo Sawada. 2018. TUFS Asian Language Parallel Corpus (TALPCo).『言語処理学会 第24回年次大会 発表論文集』, 436-439.
- Riza, Hammam, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian Language Treebank. In *Oriental COCOSDA*.
- 清水政明. 2011. 『ベトナム語』大阪大学出版会.