

アイヌ語訳『五倫名義解』 Universal Dependencies 並行コーパスへの挑戦

安岡孝一*・安岡素子*

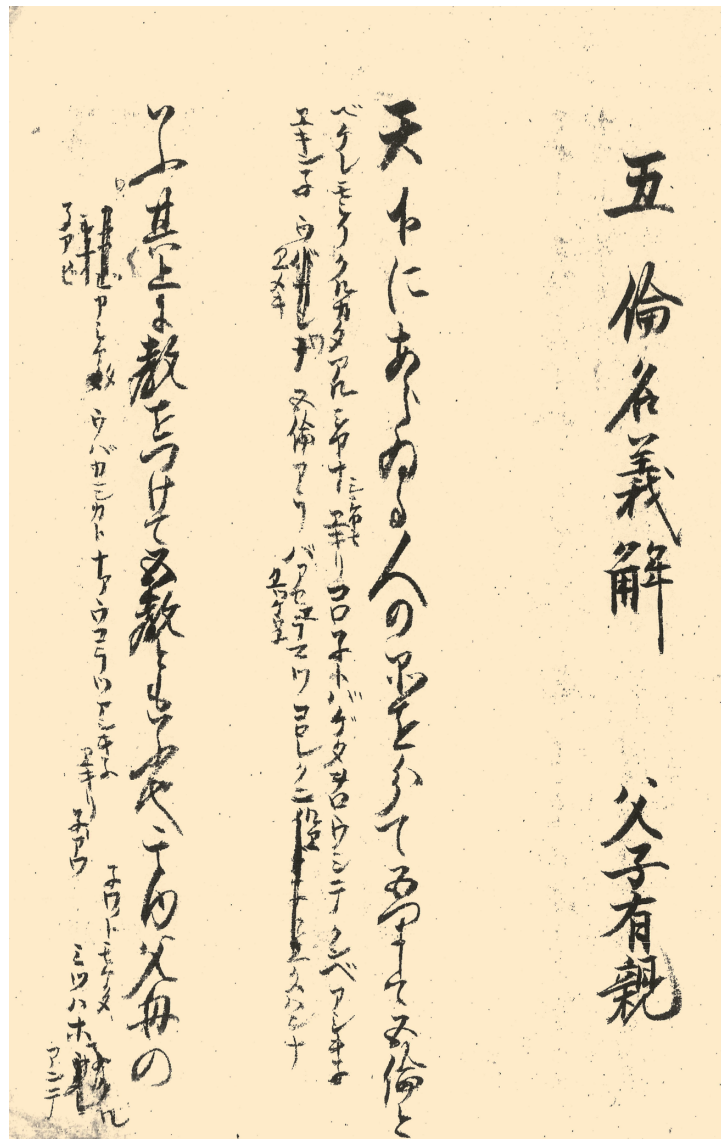


図 1: アイヌ語訳『五倫名義解』冒頭部^[1]

*京都大学人文科学研究所附属東アジア人文情報学研究センター

[1] 標津代官南摩綱紀と大通辞加賀伝蔵のアイヌ教導, 別海町郷土資料館だより, No.173 (2013年12月).

加賀家文書館(別海町)所蔵のアイヌ語訳『五倫名義解』(整理番号 K3-21)は、空谷茂潤『五倫名義解』^[2]を元に、加賀伝蔵がアイヌ語訳を施したもので、文久～慶応年間に書かれたとされている^{[1][3]}。父子有親・君臣有義・夫婦有別・長幼有序・朋友有信の五倫に加え、空谷茂潤による刊記もアイヌ語に訳されているが、アイヌ語訳には書き直しが多く(図1)、全体わずか39ページだが読みにくい。なお、加賀家文書館は『五倫名義解』をもう1冊(整理番号 K1-22)所蔵しているが、こちらにアイヌ語訳は付されていない。

北海道立図書館は、1974年10月に加賀家文書のマイクロフィルム化をおこなっており、アイヌ語訳『五倫名義解』も「加賀家文書21」として閲覧・複写可能である。秋田県公文書館は、このフィルムのコピーと紙焼きを所蔵しており、アイヌ語訳『五倫名義解』も「加賀家文書21」として閲覧・複写可能である。いずれもモノクロ撮影そのものは良好だが、元の文書が読みにくいため、やはり読みにくさに違いはない。なお「加賀家文書22」に、もう1点『五倫名義解』が収録されているが、こちらにアイヌ語訳は付されていない。

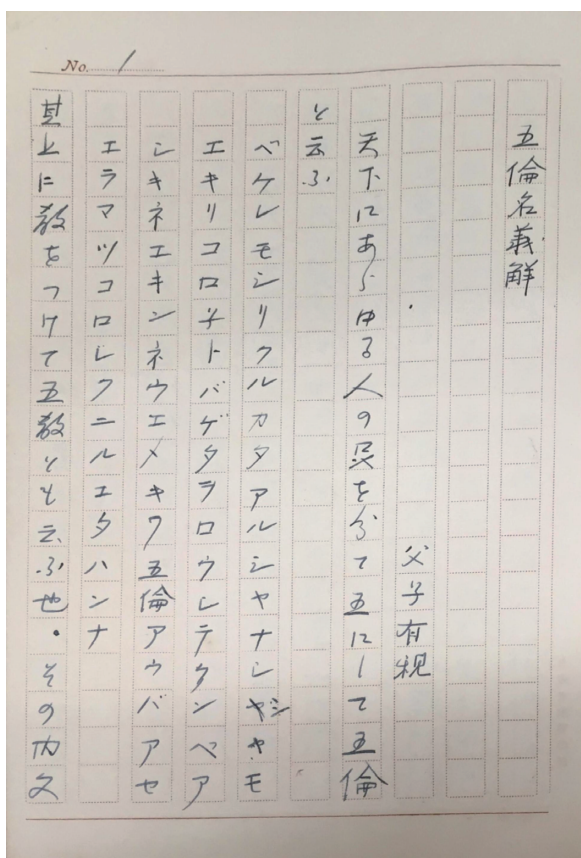


図2: 長尾元一郎寄贈『五倫名義解』後半冒頭部

^[2]空谷茂潤『五倫名義解』には、われわれが知る限り、此君園(安政2年)版と宗谷御用所(安政5年)版があるが、内容はほぼ同一である。

^[3]深澤美香: 加賀家文書のアイヌ語資料と加賀伝蔵, 千葉大学大学院人文社会科学研究所研究プロジェクト報告書, 第274集『アイヌ語の文献学的研究(1)』(2014年2月), pp.21-48.

函館市中央図書館所蔵のアイヌ語訳『五倫名義解』(所蔵番号 181140341-7)は、E東京原稿用紙にペン書きの書写原稿(図2)であり、長尾元一郎が寄贈(1939年12月21日)したものである。前半で「加賀家文書22」を、後半で「加賀家文書21」を書写しており、いずれも父子有親・君臣有義・夫婦有別・長幼有序・朋友有信の五倫に加え、空谷茂潤による刊記も含んでいる。奥付は「空谷茂潤誌」「加賀屋傳蔵訳」「五倫名義解」「アイヌ語訳」と書かれていて、日付はない。読みやすくはあるものの、ところどころ誤脱があり、たとえば図2では、図1の「バアセエラマツ」横の「エツケウエ」が脱落している。

では「バアセエラマツ」のどこに「エツケウエ」を挿入すべきなのか。図1の「エツケウエ」は「イツケウエ」に重ね書きした形跡があり、どうも ikkewe のようである。一方「バアセエラマツ」は、「ラマツ」を ramat と見るなら、「バアセエ」が páse に当たりそうだ。ikkewe には、直前の名詞の「最上級」を表す用法がある(らしい)ので、だとすると動詞 páse ではなく、名詞 ramat の直後に置くべき^[4]ということになる。つまり「バアセエラマツ」の直後に「エツケウエ」を挿入すべきである。

このような作業をアイヌ語訳『五倫名義解』に対しておこなうには、単なる全文テキストでは不十分で、少なくとも形態素解析の力を借りる必要がある。できれば係り受け解析もほしい。アイヌ語の形態素解析は、複数の解析ツールが開発されている^{[5][6][7]}ものの、しかしながら係り受け解析は今のところ、アイヌ語 Universal Dependencies^{[8][9]}による解析ツール^[7]に頼るしかない。

ならば、冒頭の父子有親だけでも、近代日本語・アイヌ語 Universal Dependencies 並行コーパスを試作してみよう。そう決心して、われわれは作業を始めた。近代日本語 UD に関しては、SuPar-UniDic^[10]の近代文語 UniDic モードで仮コーパスを作成し、deplacy^[11]の UD エディター(日本語向け改造版)で編集をおこなった。アイヌ語 UD に関しては、esupar^[12]のアイヌ語 DeBERTa モードで仮コーパスを作成し、deplacy の UD エディター(アイヌ語向け改造版)で編集をおこなった。作業結果^[13]を、次ページ以降に見開き並行コーパスの形で示す。なお、UD における各タグの意味については、付録を参照されたい。

^[4]他の用例としては、朋友有信に「ラマツイツケウエ子ルエタバン」の一節があり、「本意とすべし」のアイヌ語訳に当てられている。

^[5]Michal Ptaszynski, Yoshio Momouchi: POST-AL: Part-of-Speech Tagger for Ainu Language, 言語処理学会第18回年次大会発表論文集(2012年3月), pp.763-766.

^[6]Karol Nowakowski, Michal Ptaszynski, and Fumito Masui: MiNgMatch: A Fast N-gram Model for Word Segmentation of the Ainu Language, Information, Vol.10, No.10 (October 2019), 317.

^[7]安岡孝一: ローマ字・カタカナ・キリル文字併用アイヌ語 RoBERTa・DeBERTa モデルの開発, 情報処理学会研究報告, Vol.2023-CH-131 『人文科学とコンピュータ』, No.7 (2023年2月18日), pp.1-7.

^[8]Hajime Senuma, Akiko Aizawa: Universal Dependencies for Ainu, LREC 2018: Eleventh International Conference on Language Resources and Evaluation (May 2018), pp.2354-2358.

^[9]安岡孝一: Universal Dependencies によるアイヌ語テキストコーパス, 情報処理学会研究報告, Vol.2021-CH-127 『人文科学とコンピュータ』, No.5 (2021年8月28日), pp.1-8.

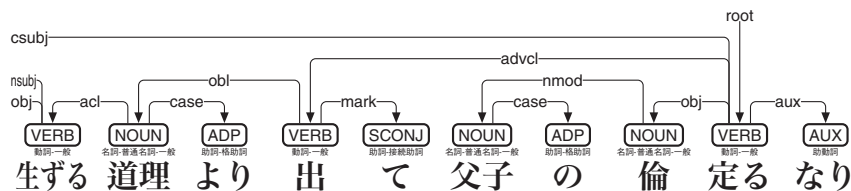
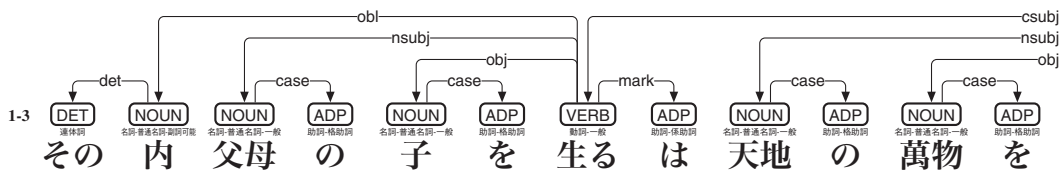
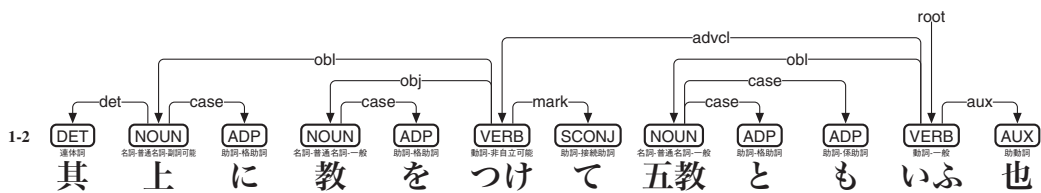
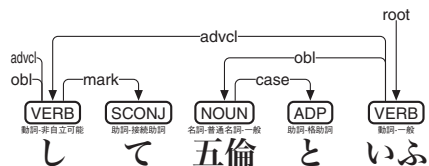
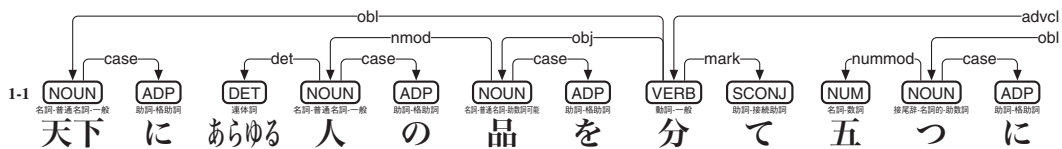
^[10]<https://github.com/KoichiYasuoka/SuPar-UniDic>

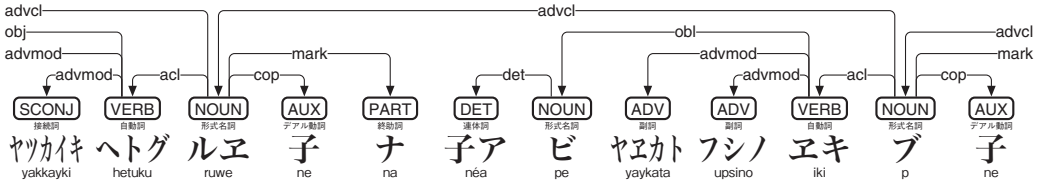
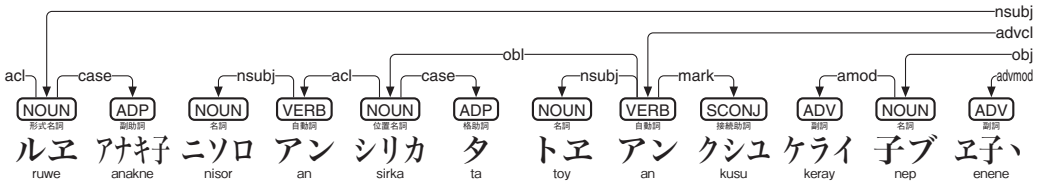
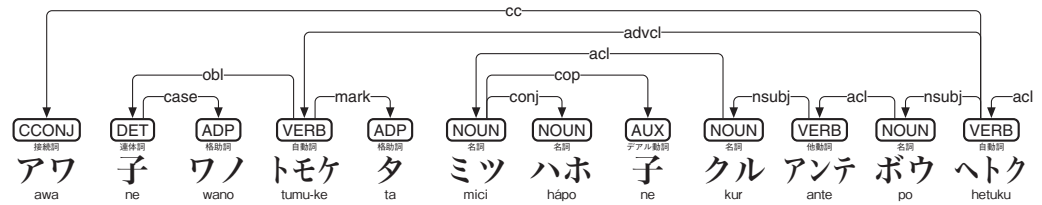
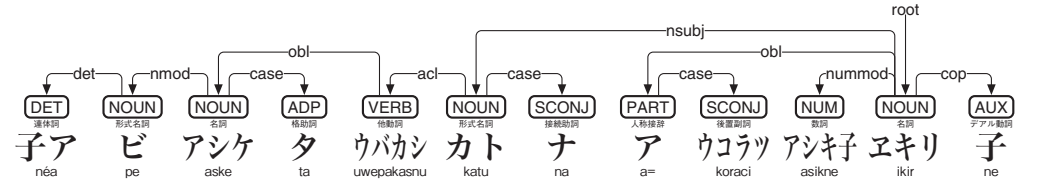
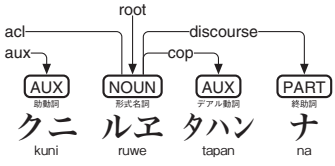
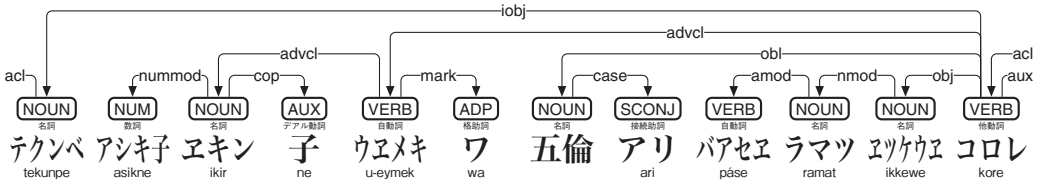
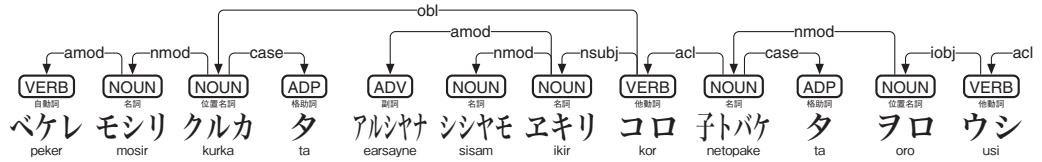
^[11]安岡孝一: Universal Dependencies にもとづく多言語係り受け可視化ツール deplacy, 人文科学とコンピュータシンポジウム「じんもんこん 2020」論文集(2020年12月), pp.95-100.

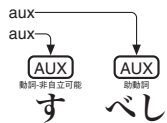
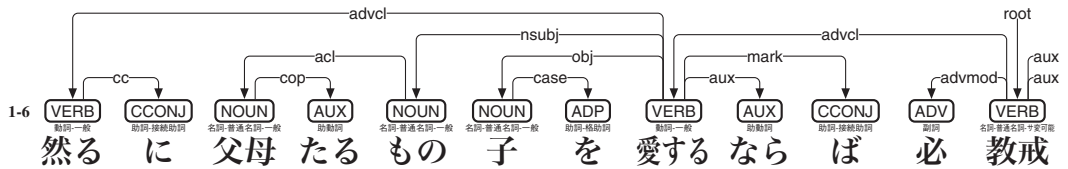
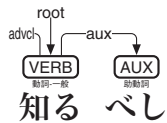
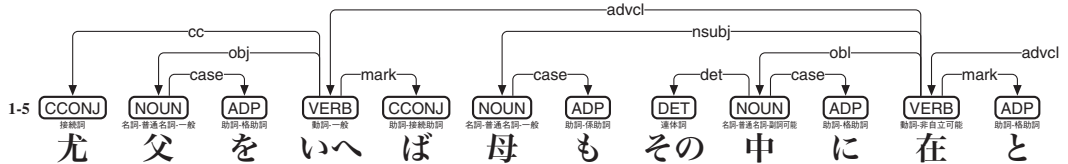
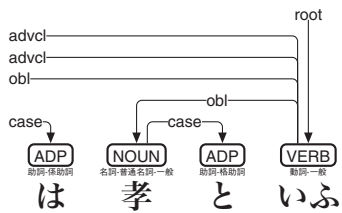
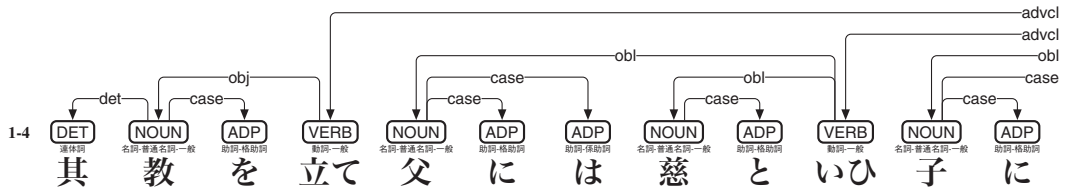
^[12]<https://github.com/KoichiYasuoka/esupar>

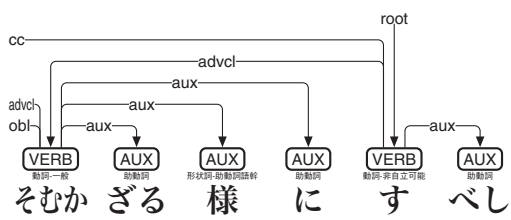
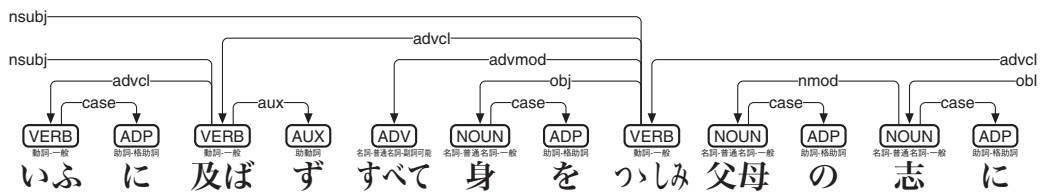
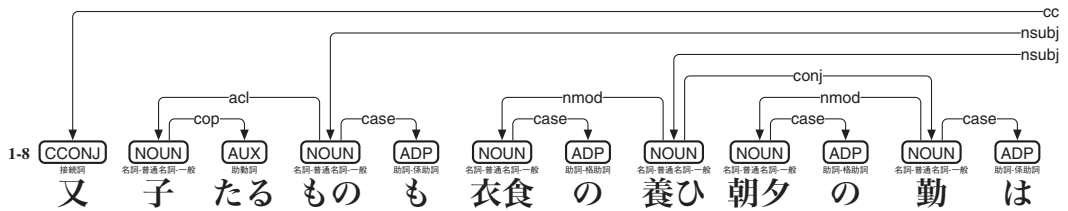
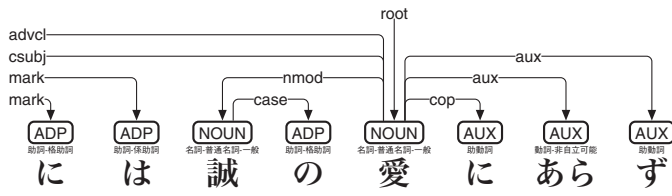
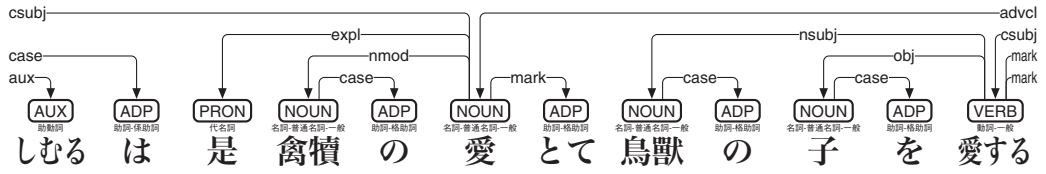
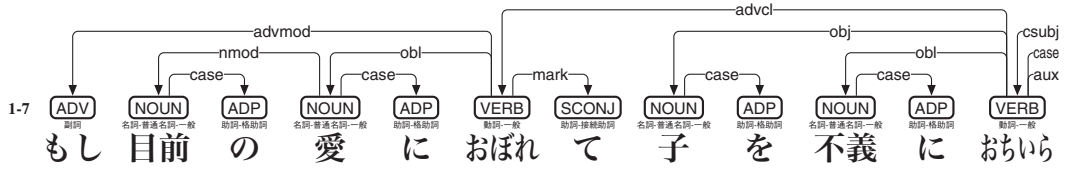
^[13]作業に用いた近代日本語・アイヌ語 UD 係り受け解析エンジンは、学際大規模情報基盤共同利用・共同研究拠点公募型共同研究『単語間に区切りのない書写言語における係り受け解析エンジンの開発』(共同研究者: 山崎直樹・二階堂善弘・師茂樹・鈴木慎吾・守岡知彦・Christian Wittern・池田巧・藤田一乗)の成果である。また、近代日本語・アイヌ語 UD エディターと並行コーパス管理システムの開発、およびそれらを用いたコーパス作成作業は、文部科学省『AI等の活用を推進する研究データエコシステム構築事業』の支援を受けている。

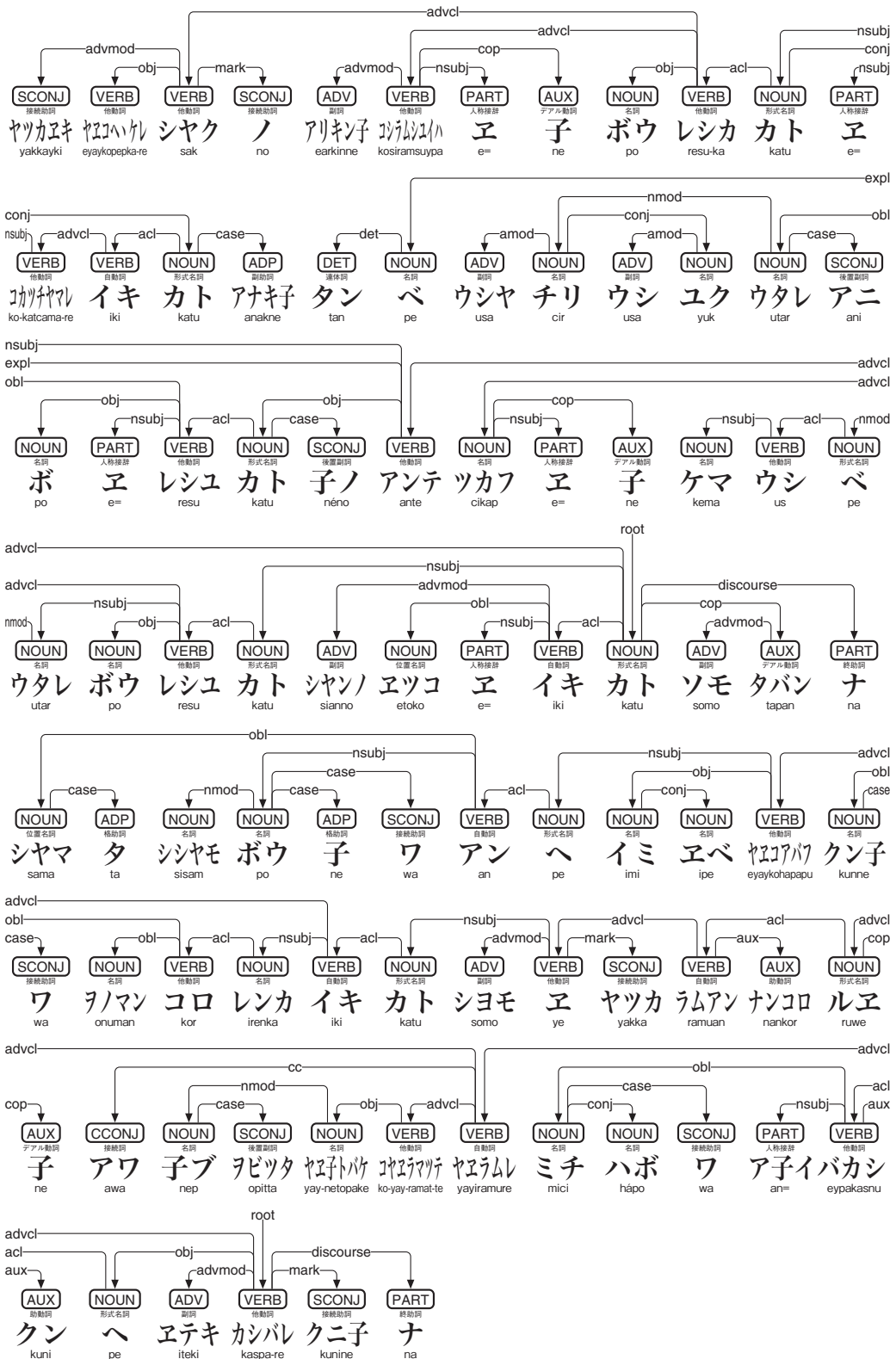
五倫名義解 父子有親

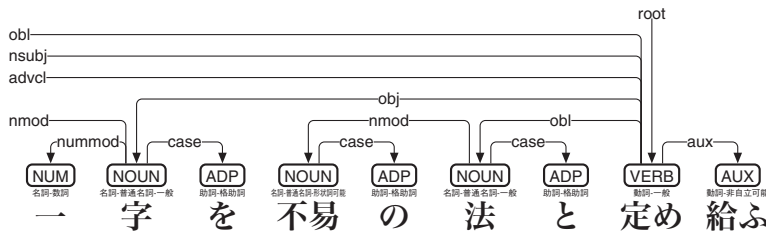
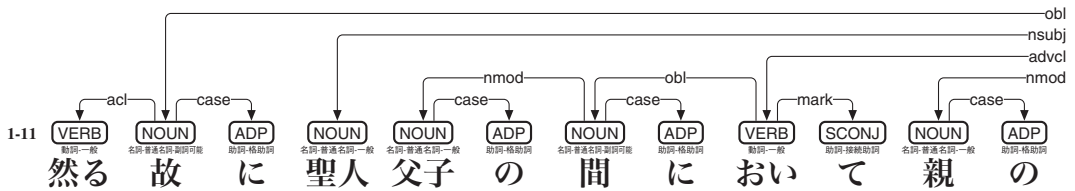
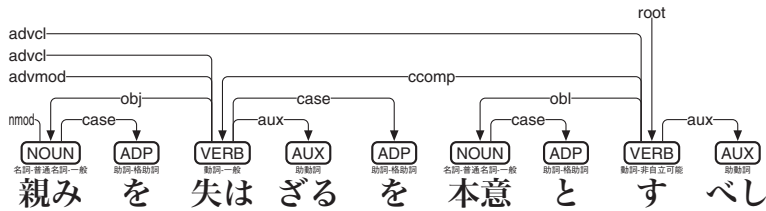
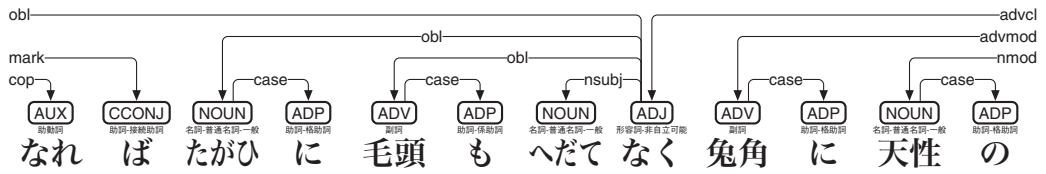
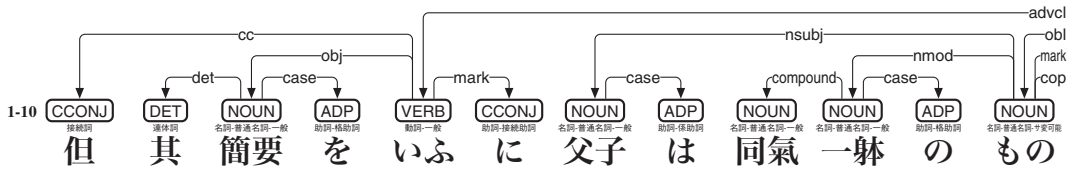
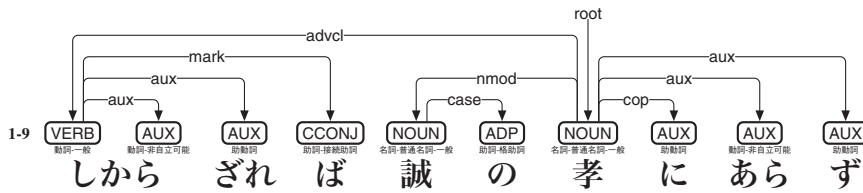


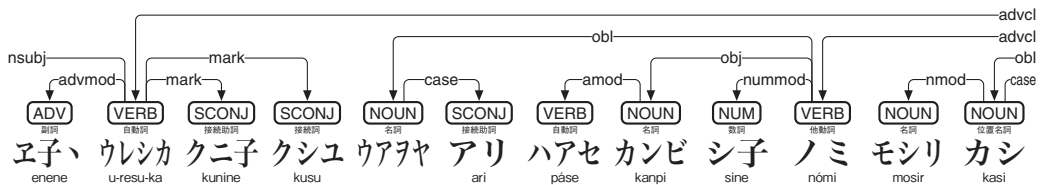
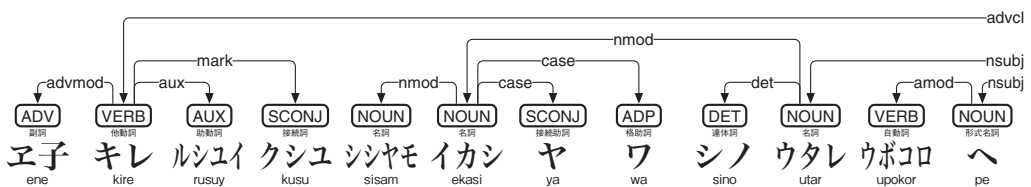
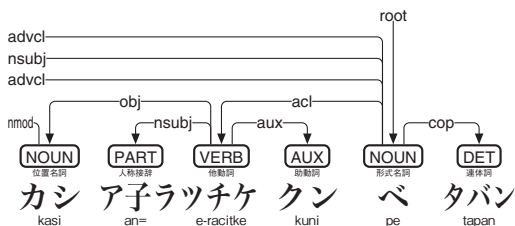
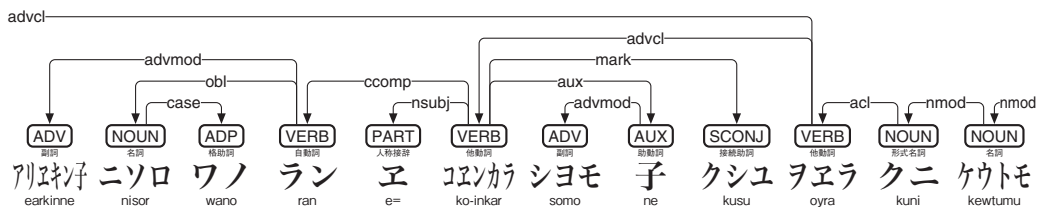
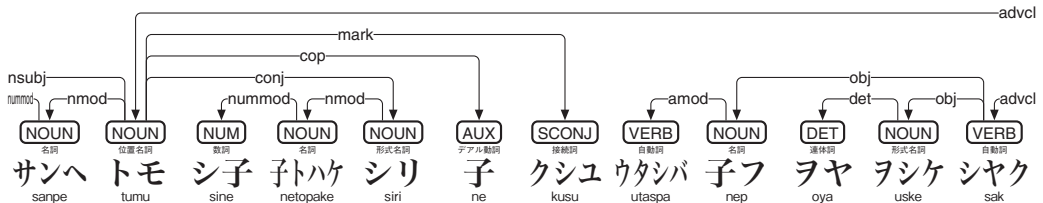
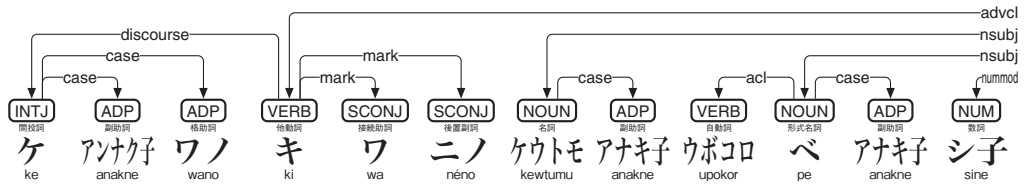
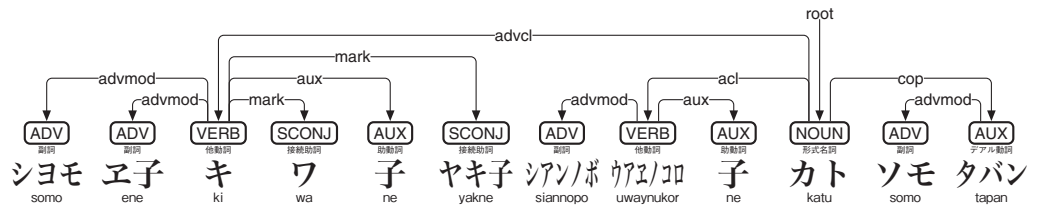


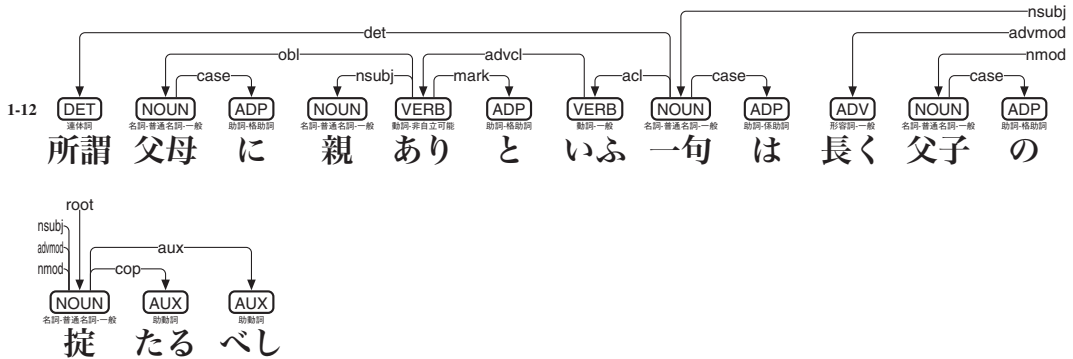










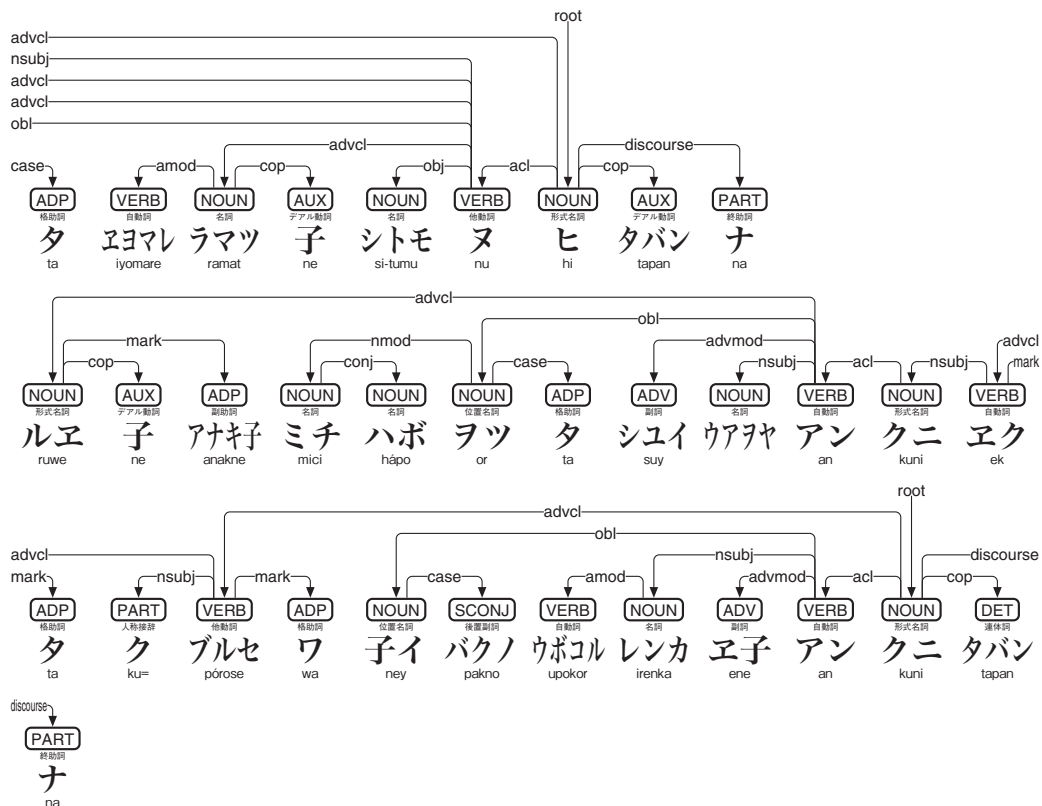


作業は困難を極めた。父子有親はわずか12対の文から構成されるが、コーパス作成作業だけで延べ100時間を要した。近代日本語UDに関しては崩し字さえ読めれば、あとはSuPar-UniDicの形態素解析・係り受け解析結果を、多少編集すれば大丈夫だった。困難を極めたのはアイヌ語UDだった。

父子有親におけるアイヌ語カタカナ表記には、小書きのカタカナが使われていない。拗音も促音も小書きにしない上に、末子音が母音を伴っており、たとえばsisamは「シシヤモ」と表記される(1-1・1-8・1-11)^[14]。しかも、表記にかなり揺れがあり、たとえばanakneは「アナキ子」だったり「アンナク子」だったりする(1-10)。hetukuは「セトク」に「ヘトク」が重ね書きされている(1-3)。アイヌ語DeBERTa内蔵のカタカナ・ローマ字変換では、正直なところ歯が立たない。

そこで、アイヌ語UDに関しては、作業を4段階に分けることにした。まずは元のカタカナ文を、esuparで解析して、仮のローマ字文(単語切り結果)を得る。仮のローマ字文を、適切なローマ字のアイヌ語の文に手作業で直し、再びesuparで解析する。ローマ字での形態素解析・係り受け解析結果を、アイヌ語UDエディターで編集する。最後に、アイヌ語UDの第2フィールド(FORM)をカタカナに戻して、アイヌ語UDコーパスを完成する。

4段階の作業のうち、2段階目において適切なローマ字を決めるのが、われわれには、かなり難しかった。1-1では「アルシヤナ」にearsayneを当てているが、これは今でも自信がない。1-2「ウバカシ」にはuwepakasnuを、1-4「エバカシアン」にはepakasnu=anを、



1-6 「イバカシイチャコク」には epakasnu ecakoko を、1-8 「ア子イバカシクンへ」には an=eypakasnu kuni pe を、それぞれ当ててみたものの、どうしても疑義^[15]が残る。1-8 「コヤエラマツテ」には ko-yay-ramat-te を、1-10 「ア子ラツチケ」には an=e-racitke を当てたが、やはり自信がない。1-11・1-12 「ウアヲヤ」は、結局わからなかった。父子有親の「親」のアイヌ語訳で、名詞だとは思うのだが、適切なローマ字を決めきれなかった。

1-4 の「アチャ子ワエシヌ、カクルボウ子ワウワエノコロクル子レコレ」は、UDでの記述が難しい構造を持っている。この部分は「áca ne wa ésir núnuke kur a=rékore」と「po ne wa uwaynukor kur a=rékore」という2つの文の両方から a=rékore を除き、代わりに ne rékore を1つだけ末尾に置いた文だと理解できる。つまり、動詞 rékore を2つの文で共有しており、直接目的語と間接目的語が2組ぶらさっている、ということになるが、そのような構造はUDでの記述が難しい。とりあえず前半の組を独立させて、rékore から advcl で繋いだものの、これでいいのかわからないところである。

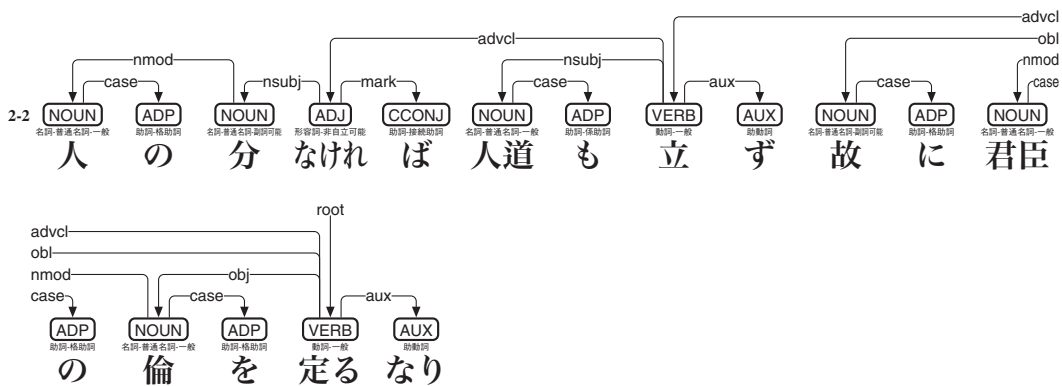
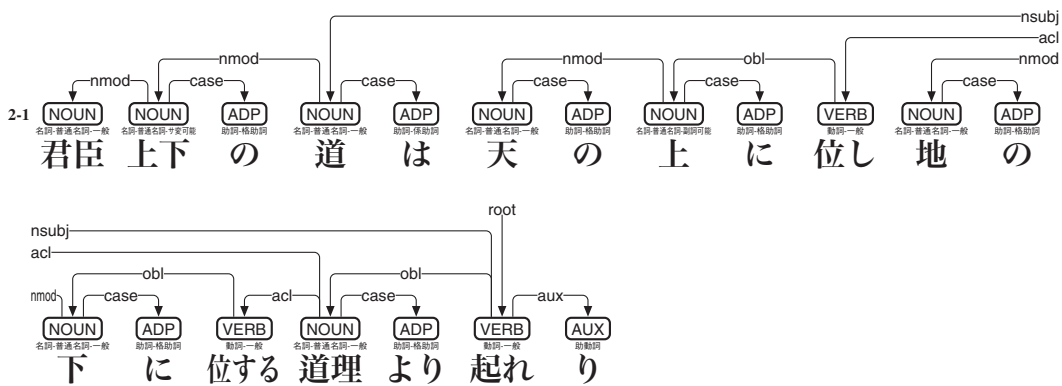
七転八倒ではあるものの、それでも何とか、12対の父子有親UD 並行コーパスを試作して、UD エディターとともに WWW で公開^[16]した。これに気を良くしたわれわれは、次の君臣有義へと作業を進めた。

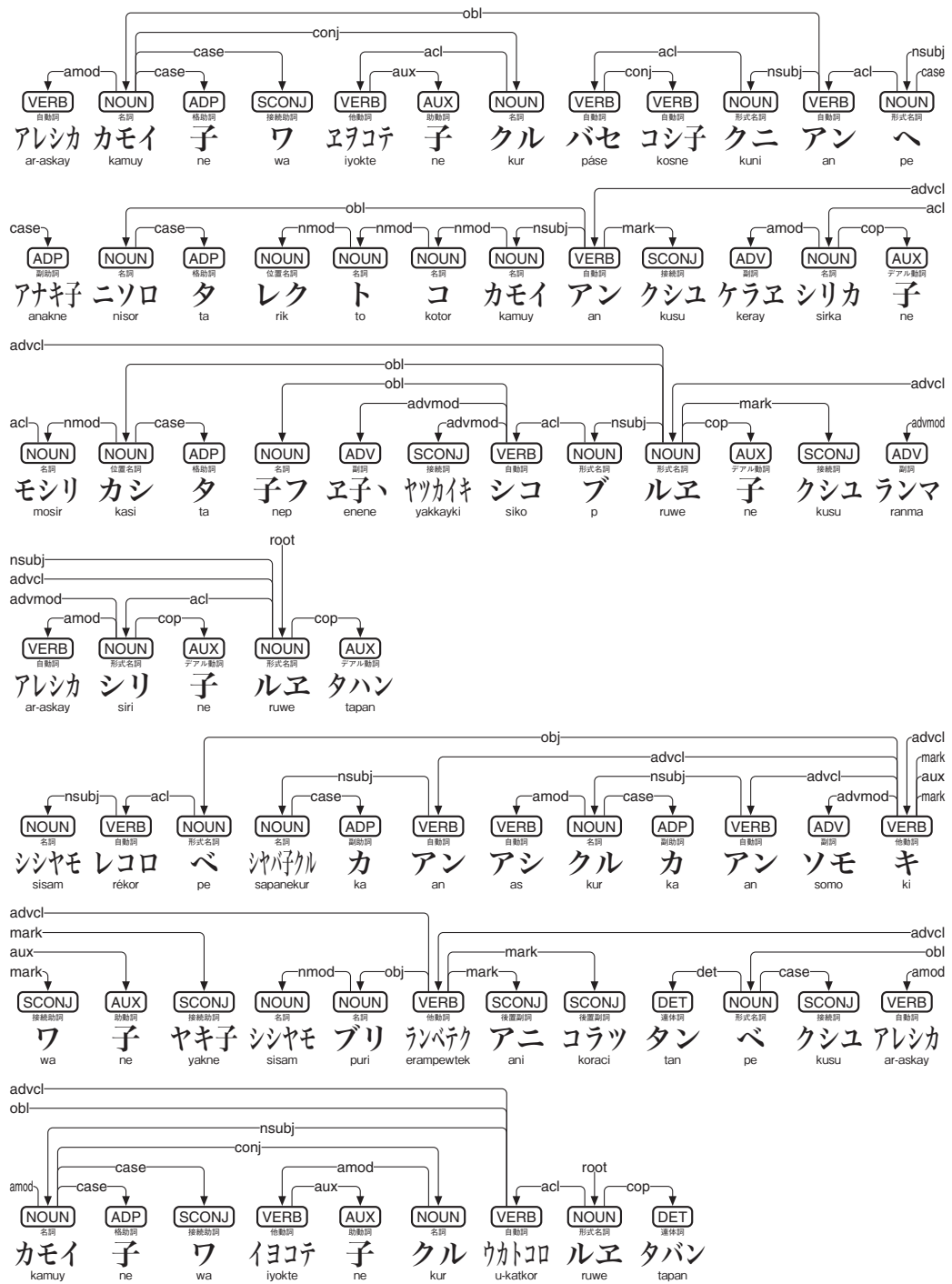
^[14] 深澤美香: 加賀家文書における表記の特徴と傾向, 千葉大学大学院人文社会科学研究所研究プロジェクト報告書, 第274集『アイヌ語の文献学的研究(1)』(2014年2月), pp.49-72.

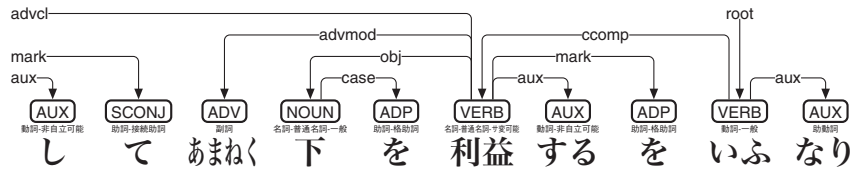
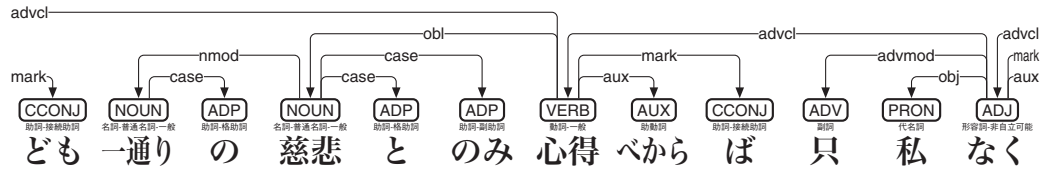
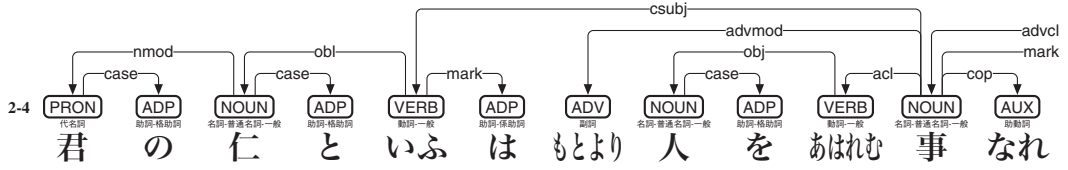
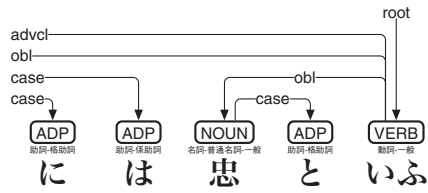
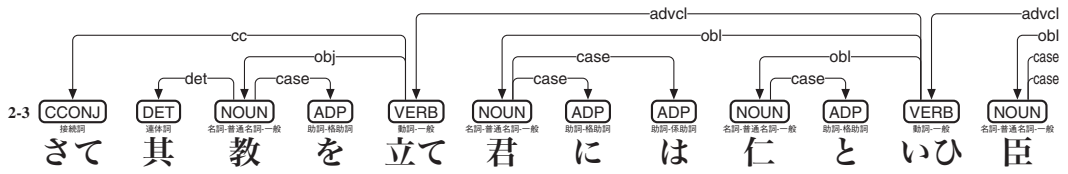
^[15] 佐藤知己: 「申渡」のアイヌ語訳文に関する一考察, 北海道立アイヌ民族文化研究センター研究紀要, 第11号(2005年3月), pp.1-45.

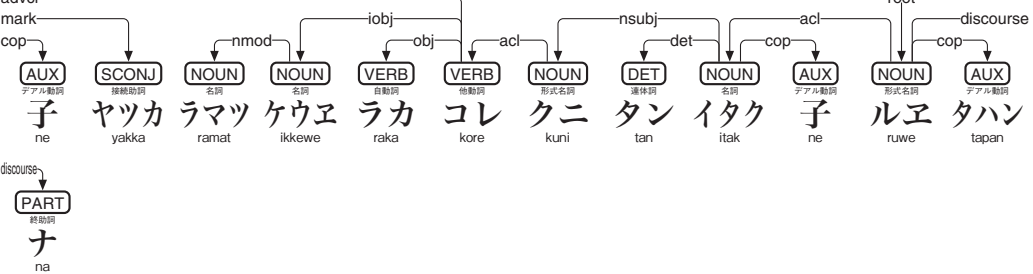
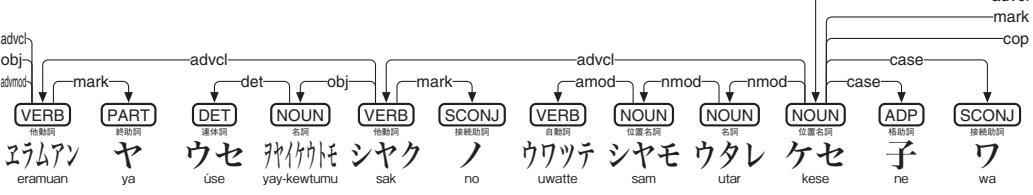
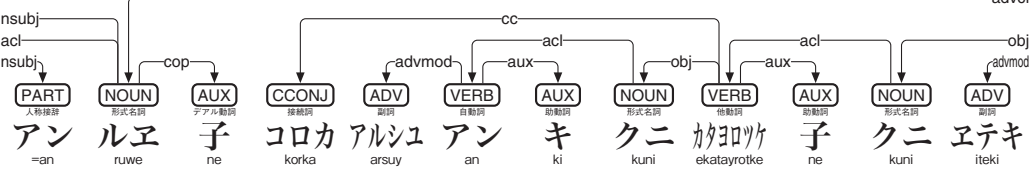
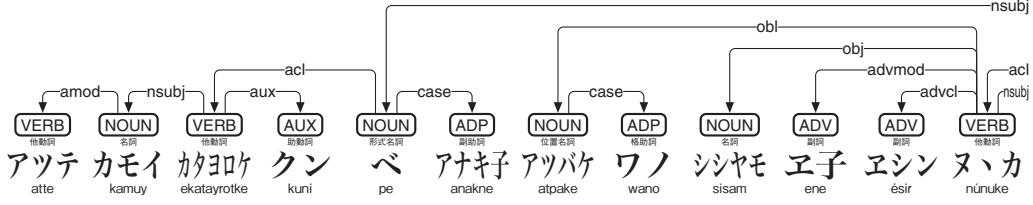
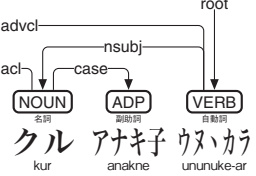
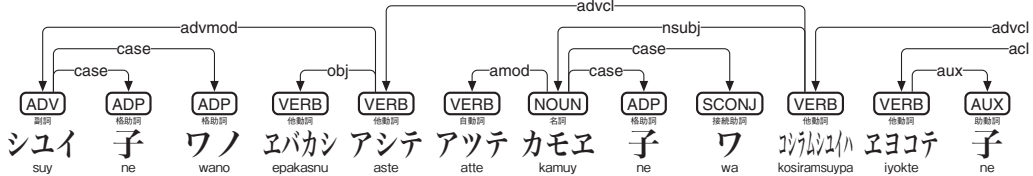
^[16] <https://koichiyasuoka.github.io/deplacy/demo/2023-07-28/>

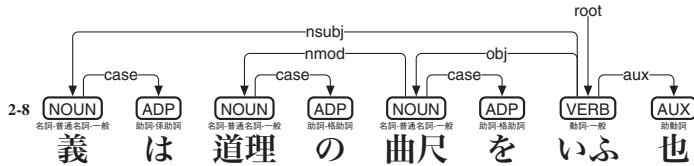
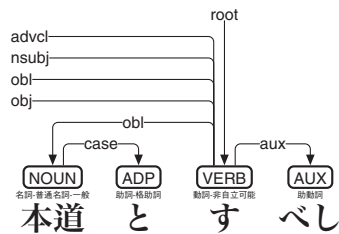
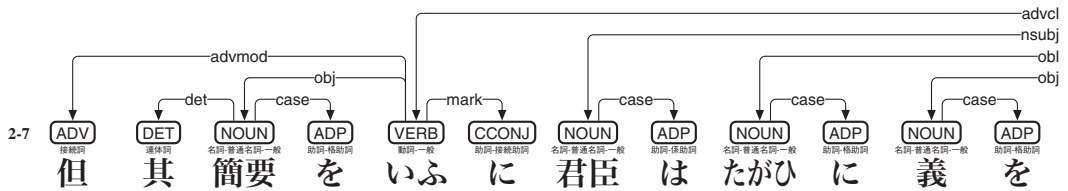
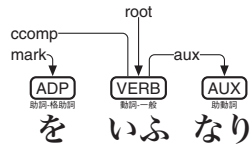
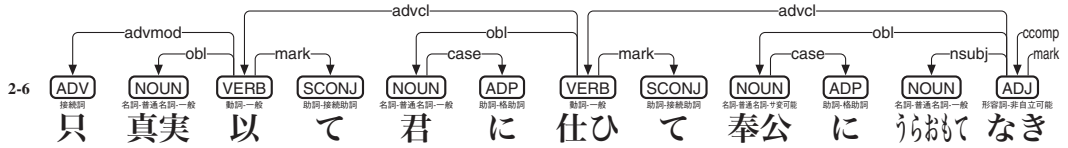
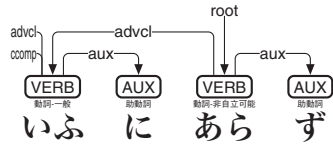
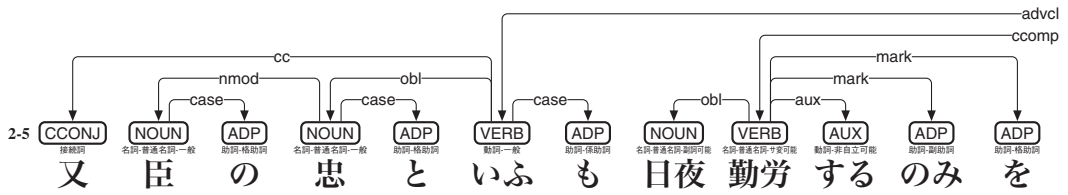
君臣有義

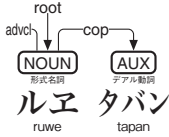
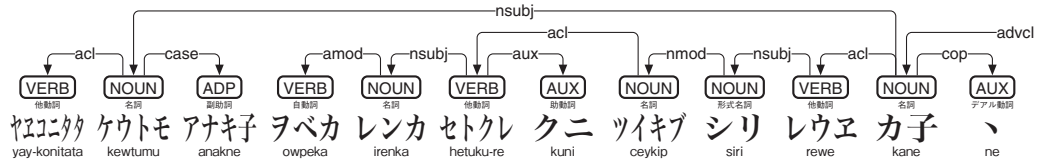
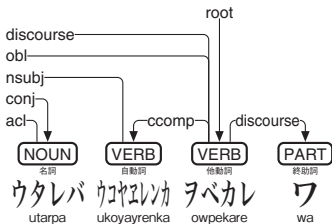
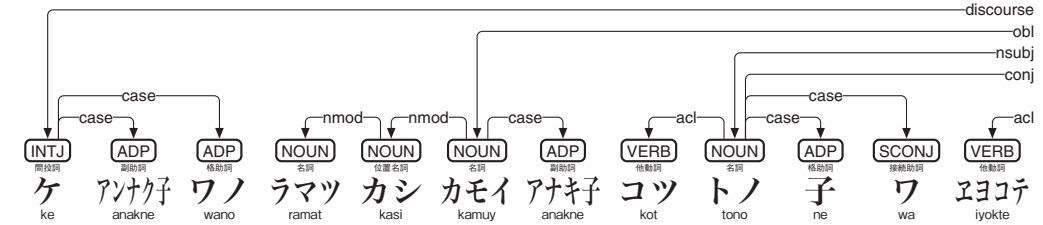
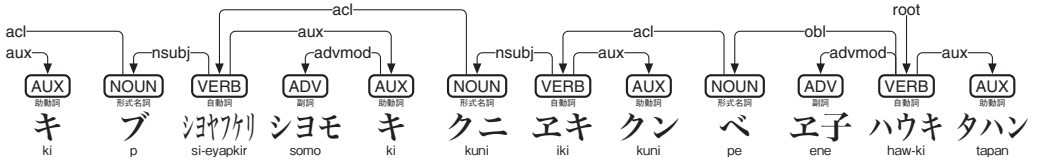
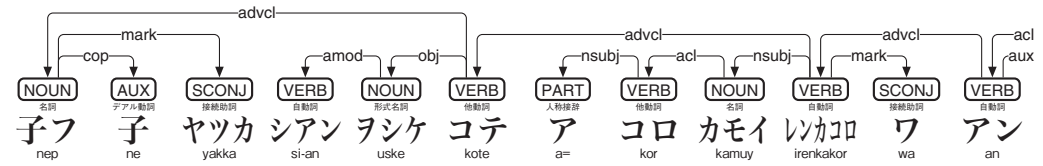
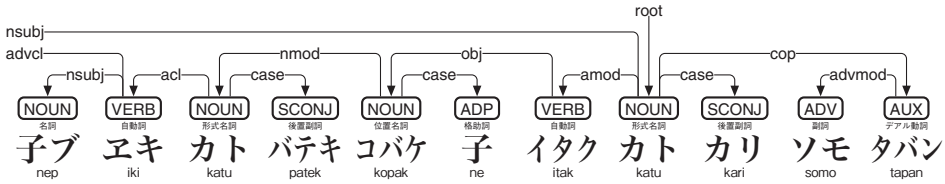
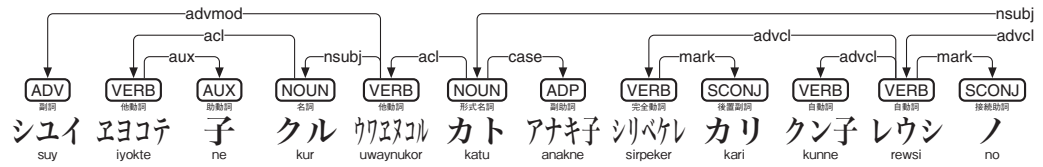


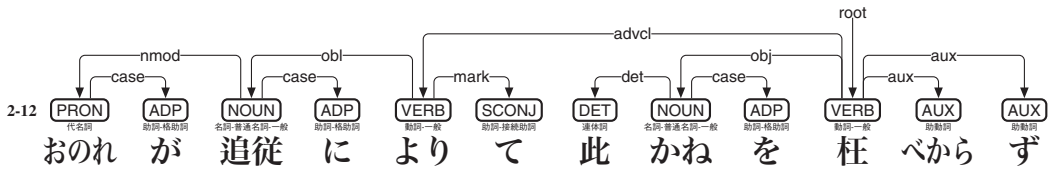
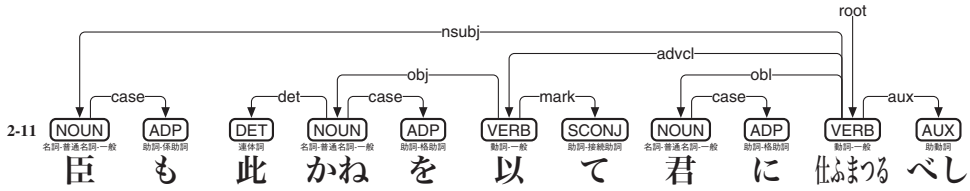
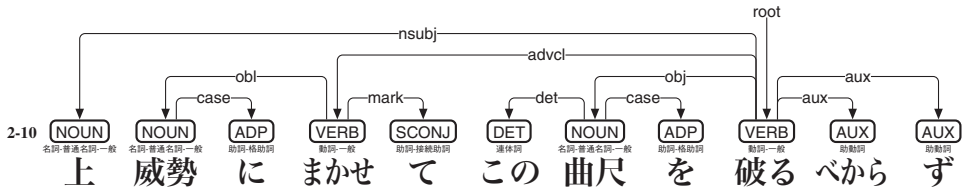
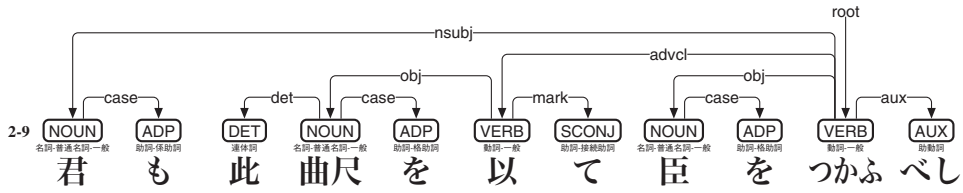


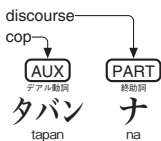
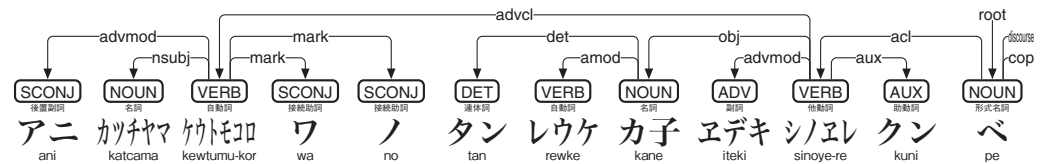
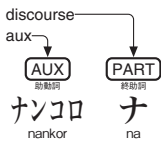
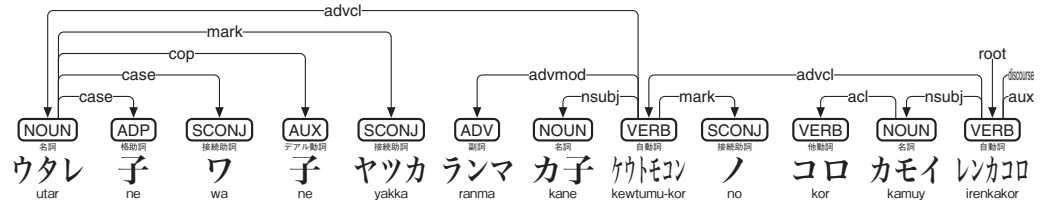
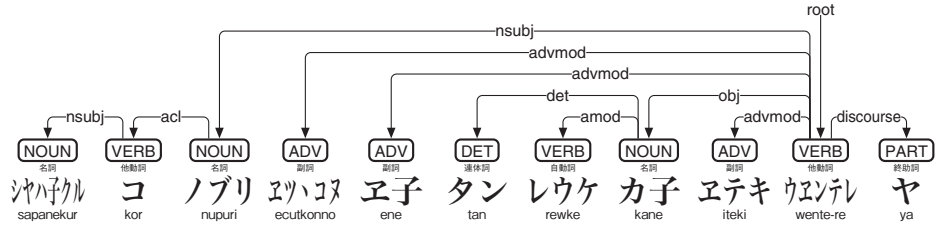
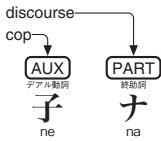
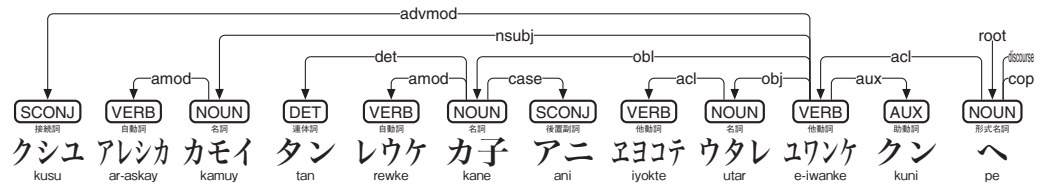


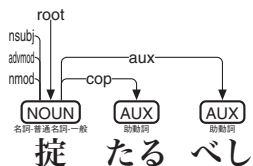
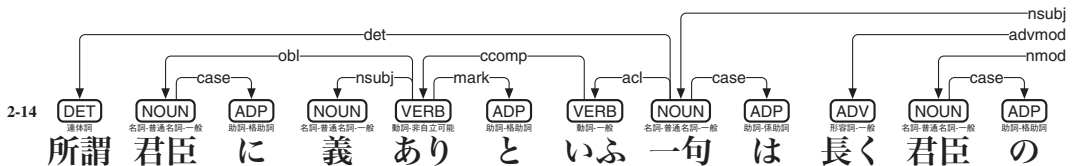
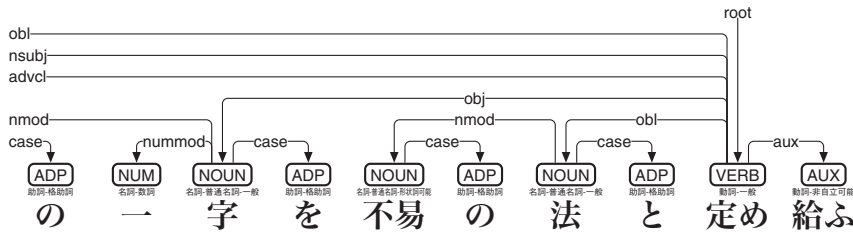
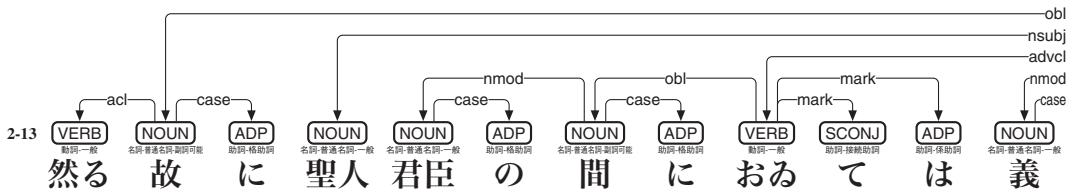






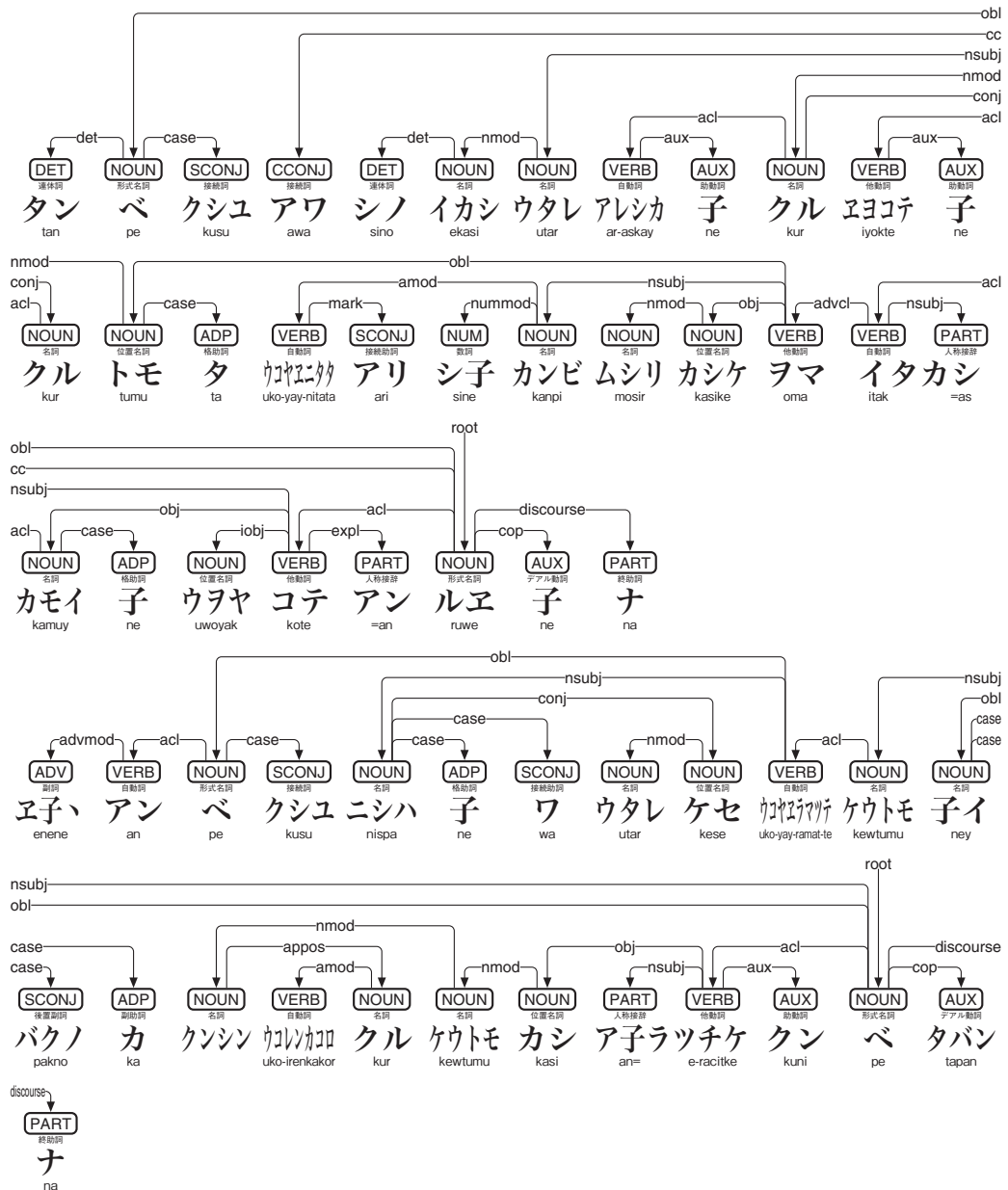






君臣有義は 14 対の文から構成されるが、コーパス作成作業に延べ 50 時間を要した。UD 並行コーパスの作成手順は、父子有親と同様である。

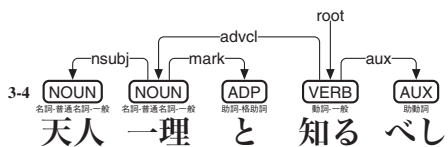
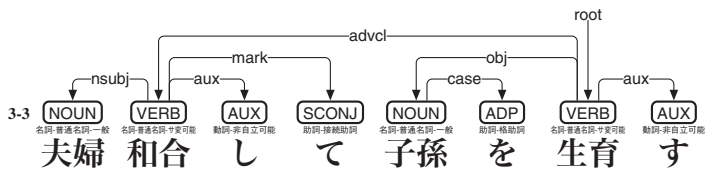
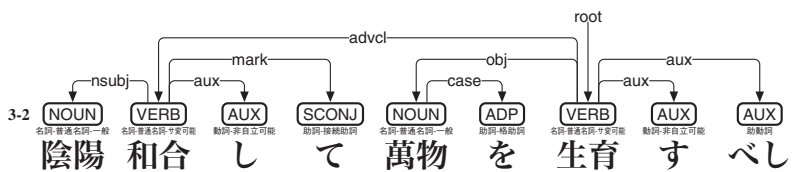
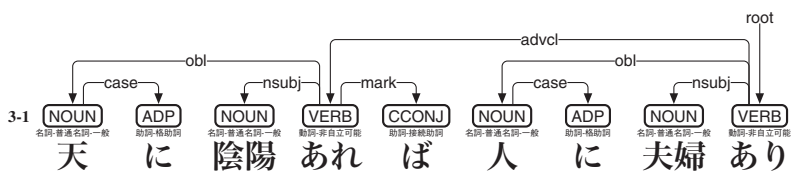
君臣有義では「君」のアイヌ語訳が、「アレスカカモイ」(2-1・2-2・2-9)、「アレスカ子クル」(2-13)、「アツテカモエ」(2-3)、「アツテカモイ」(2-4)、「アコロカモイ」(2-6)、「コロカモイ」(2-11)、「コツノ」(2-7)、「ニシハ」(2-14)と多彩である。「アレスカ」には、a=resu-ka ではなく ar-askay を当てているが、これは今でも自信がない。「臣」のアイヌ語訳も、「エヨコテ子クル」(2-1)、「エヨコテ子クル」(2-3・2-5・2-13)、「イヨコテ子クル」(2-2)、「エヨコテウタレバ」(2-7)、「ウタレケセ」(2-14)、「ウタレ」(2-11)と多彩である。また、2-14「クンシン」は「君臣」そのままである。「義」のアイヌ語訳も、「ウコヤエレンカ」(2-7)、「ヤエコニタタ」(2-8)、「ウコヤエニタタ」(2-13)、「ウコヤエラマツテ」(2-14)と多彩である。

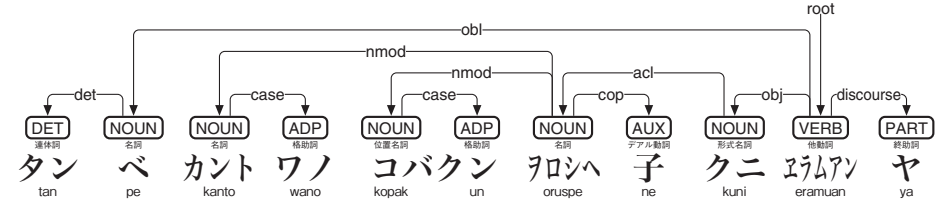
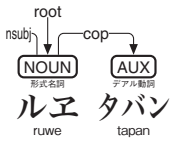
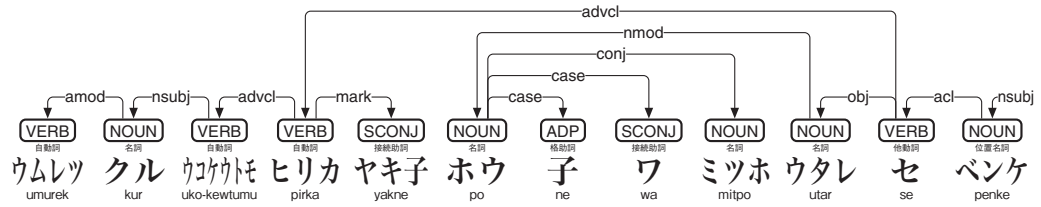
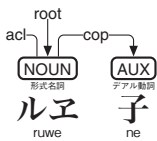
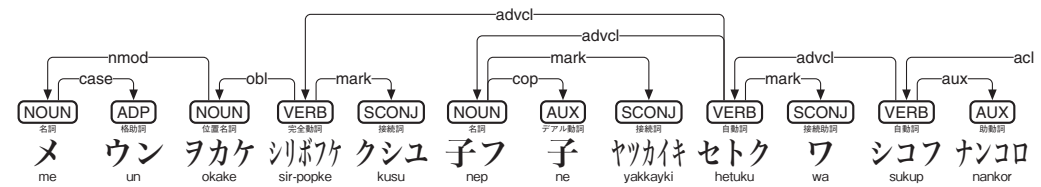
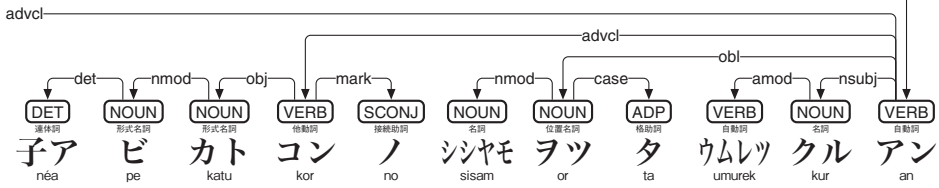
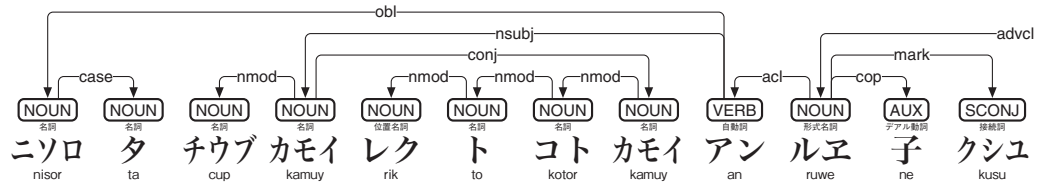


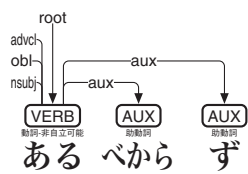
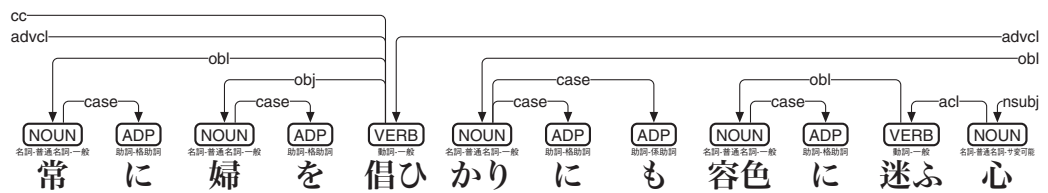
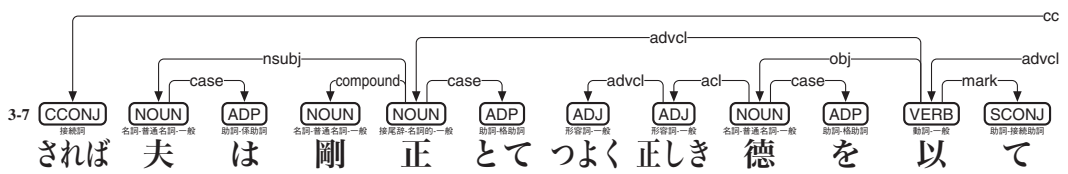
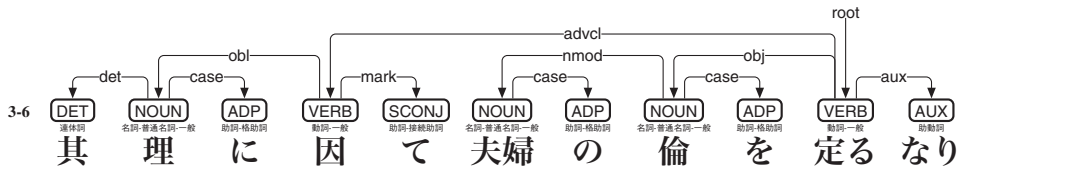
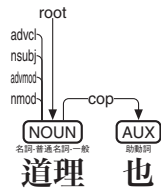
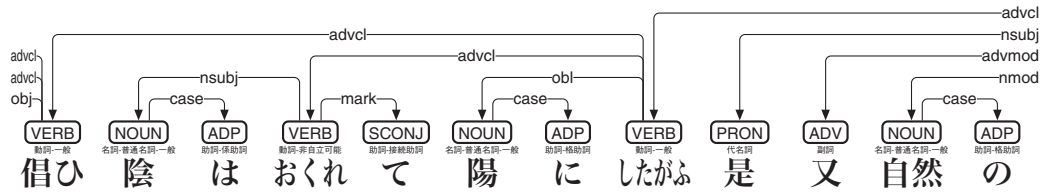
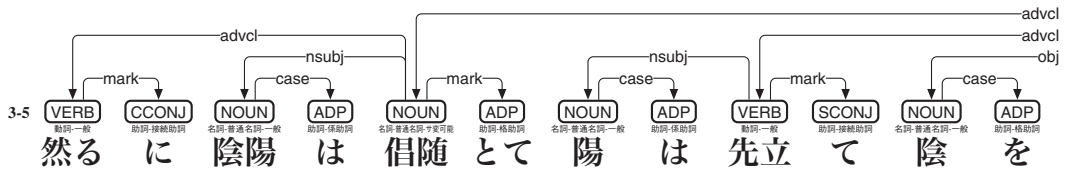
2-3 「ウヌ、カラ」には ununuke-ar を、2-6 「シヨヤフケリ」には si-eyapkir を、2-8 「セトクレ」には hetuku-re を当てたが、いずれも自信がない。2-13 「ウヲヤコテアン」に uwoyak kote=an を当ててるのは、「ヲ」と「チ」が見分けにくい^[17]こともあって、やはり自信がない。2-1 「レクトコカモイ」に rik to kotor kamuy を当ててるのは無理がある^[18]のだが、次の夫婦有別に「レクトコトカモイ」の例があり、これを借りることにした。

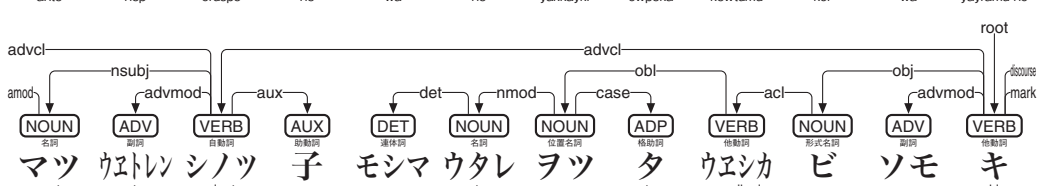
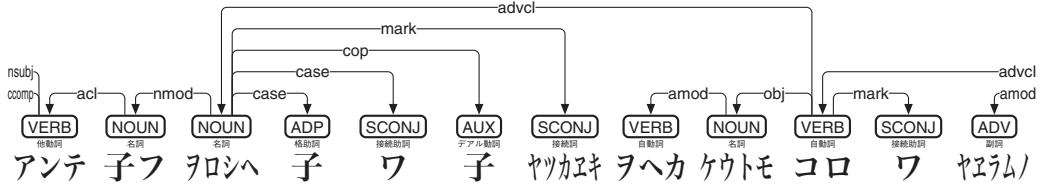
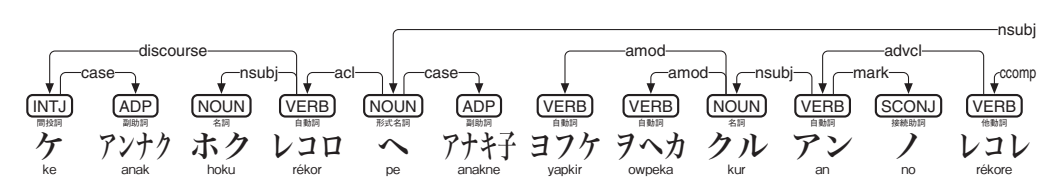
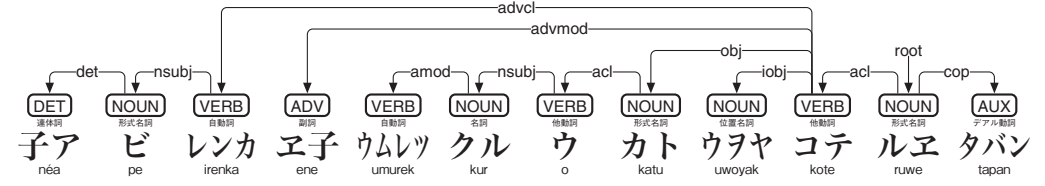
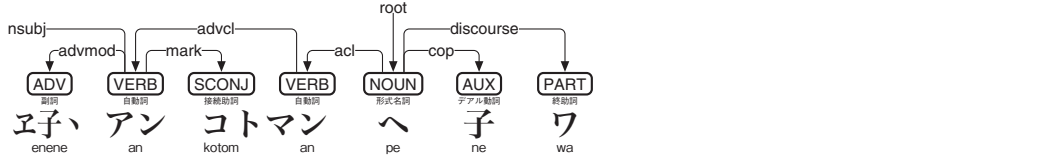
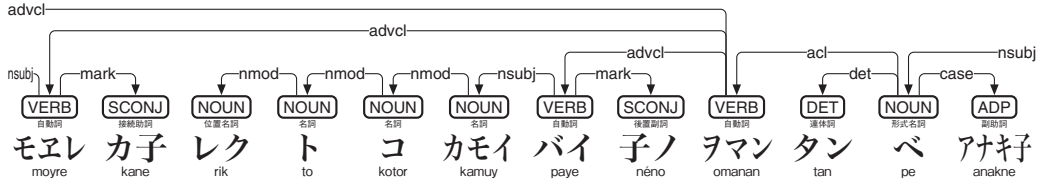
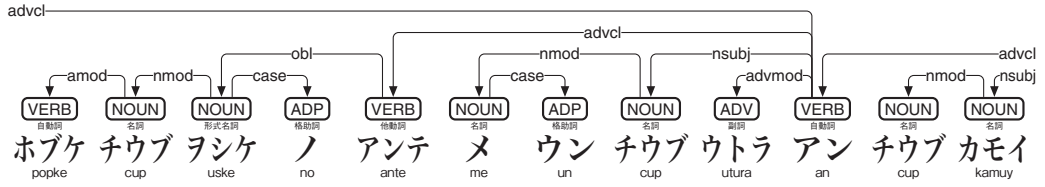
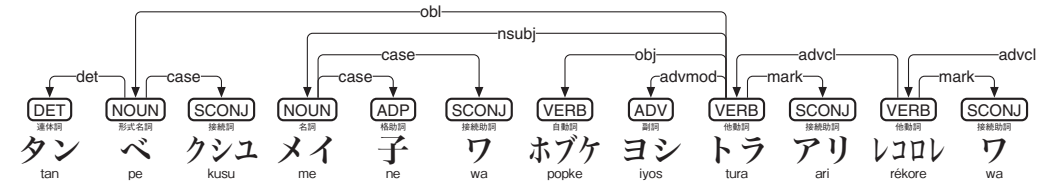
[17] 「ウチヤコテアン」ならば ocakot an だが、やはり意味が通じない気がする。
 [18] 深澤美香: 〈資料紹介〉加賀家文書「蝦夷語和解」、千葉大学大学院人文社会科学部研究科研究プロジェクト報告書、第 298 集『アイヌ語の文献学的研究 (2)』(2016 年 2 月), pp.63-101.

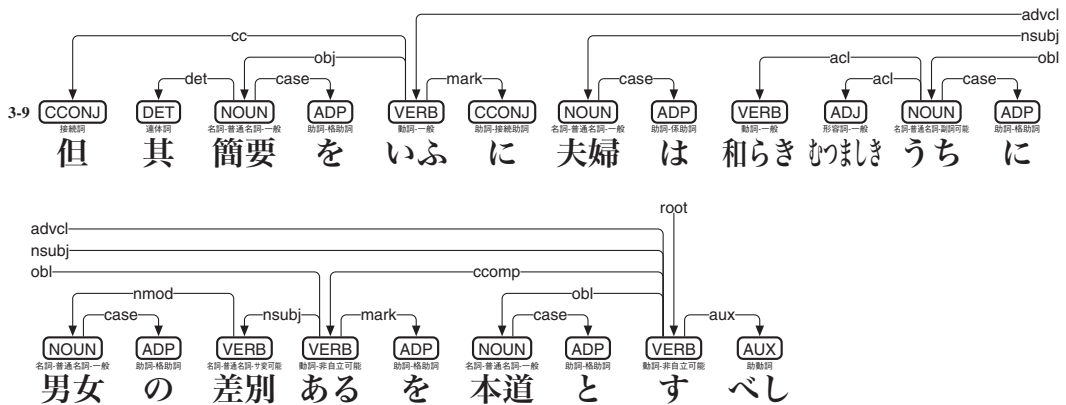
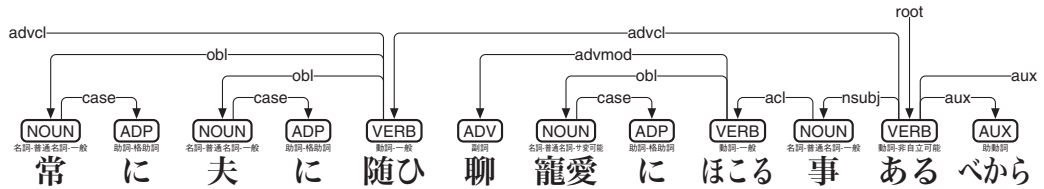
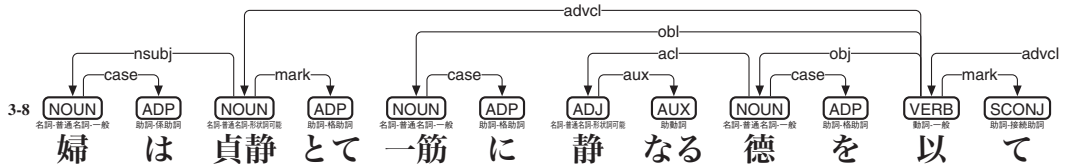
夫婦有別

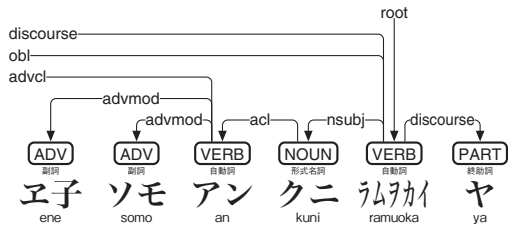
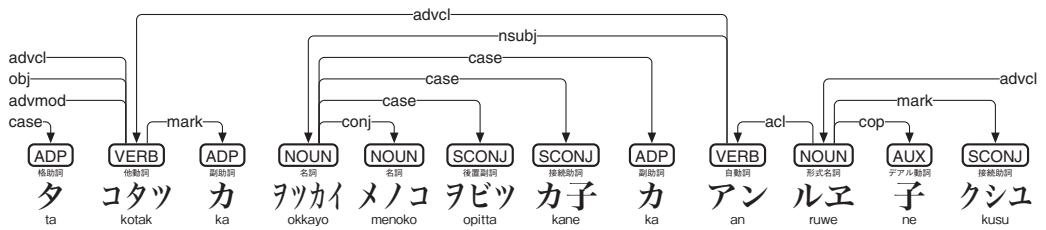
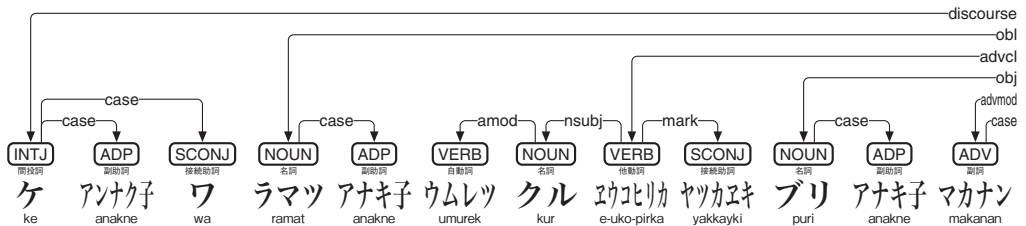
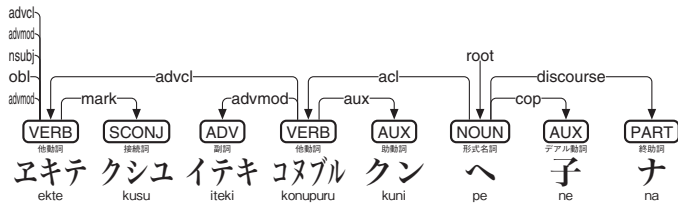
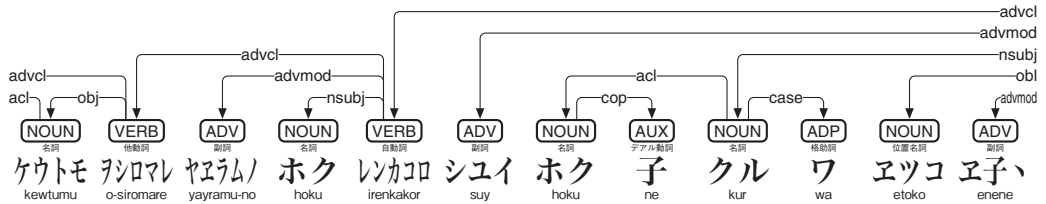
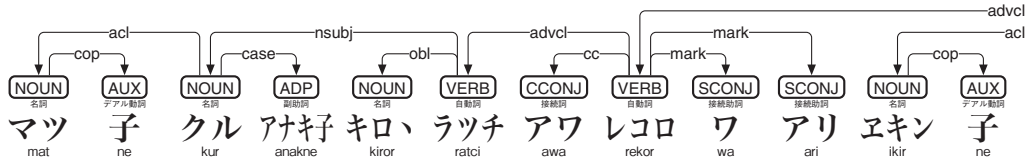
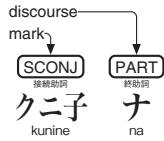


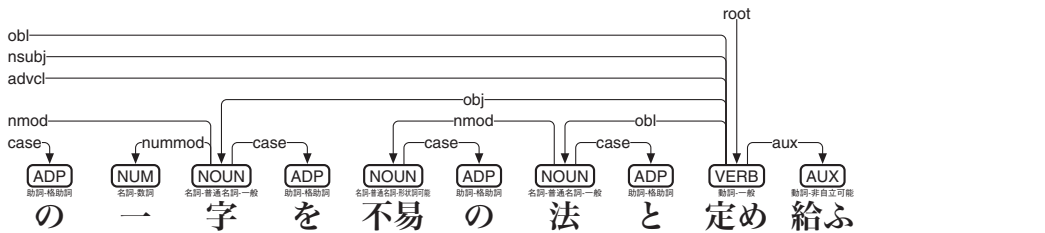
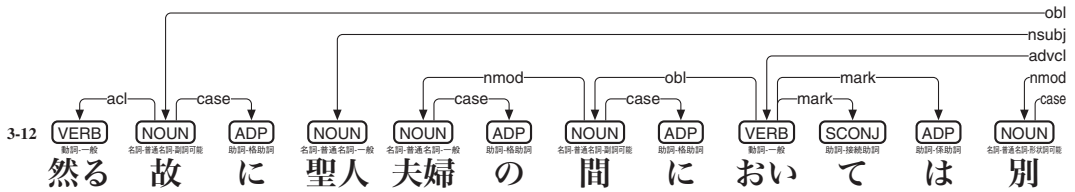
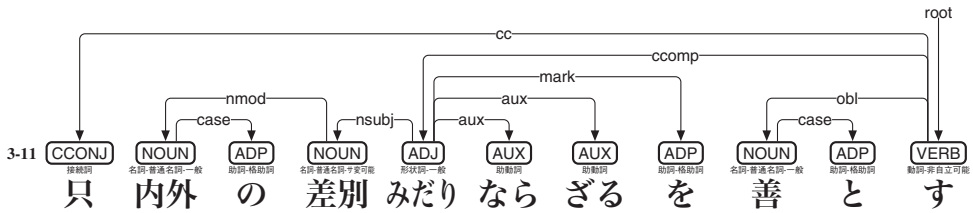
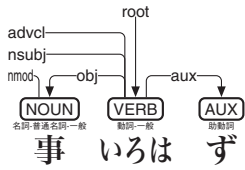
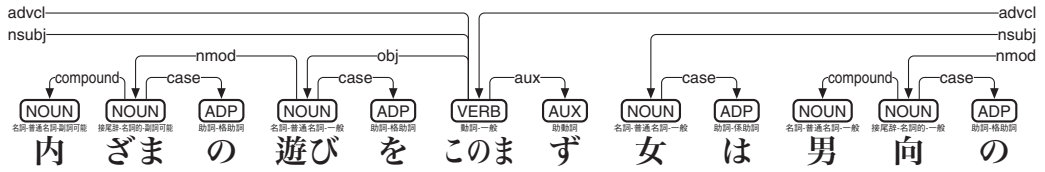
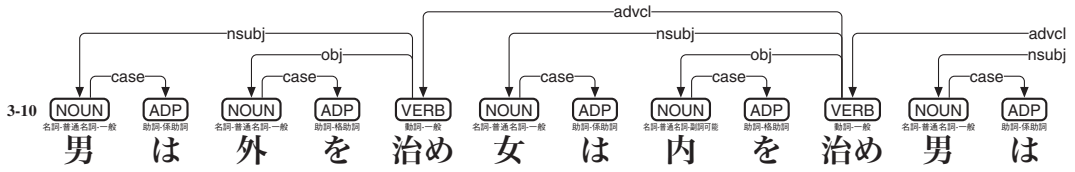


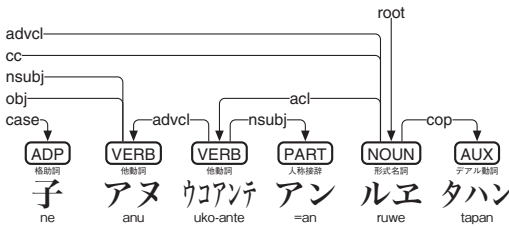
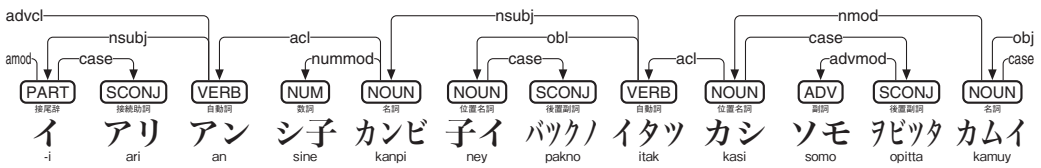
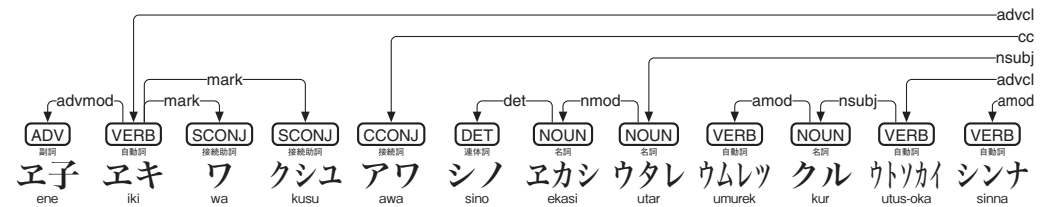
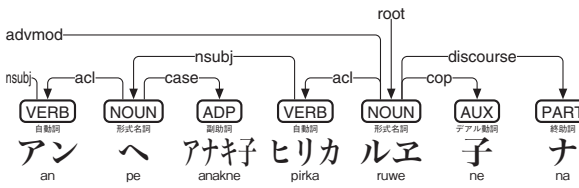
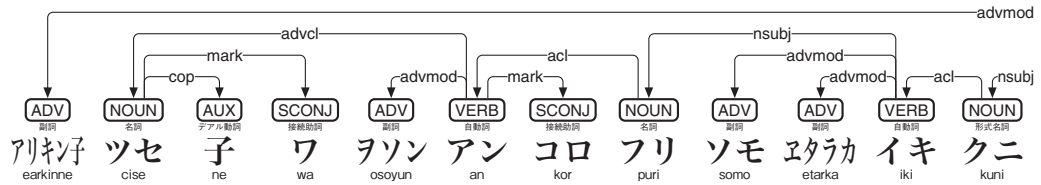
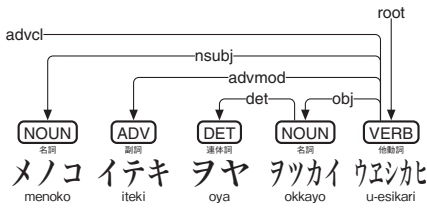
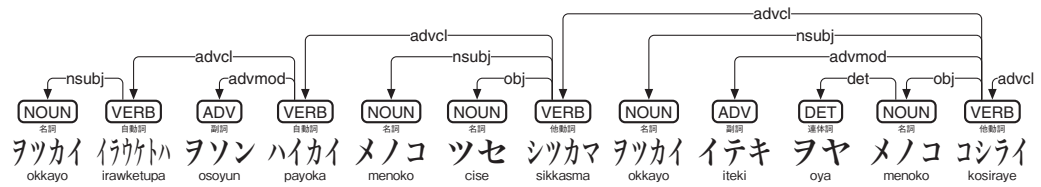


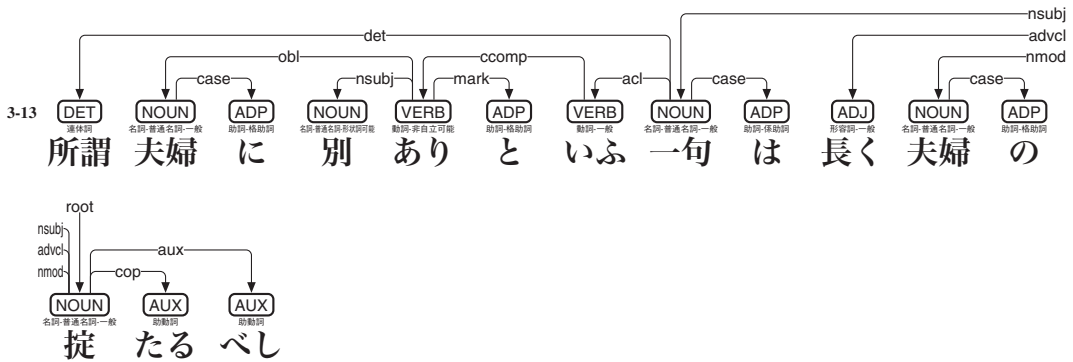












夫婦有別は13対の文から構成されるが、コーパス作成作業に延べ50時間を要した。UD並行コーパスの作成手順は、父子有親・君臣有義と同様である。なお、3-7では「ウエトレンシノツ子」と「モシマウタレ」の間に「カイ」が見えるのだが、この「カイ」がどういいうわけか筆書ではない。意味もうまく通じないことから、コーパスからは除外した。

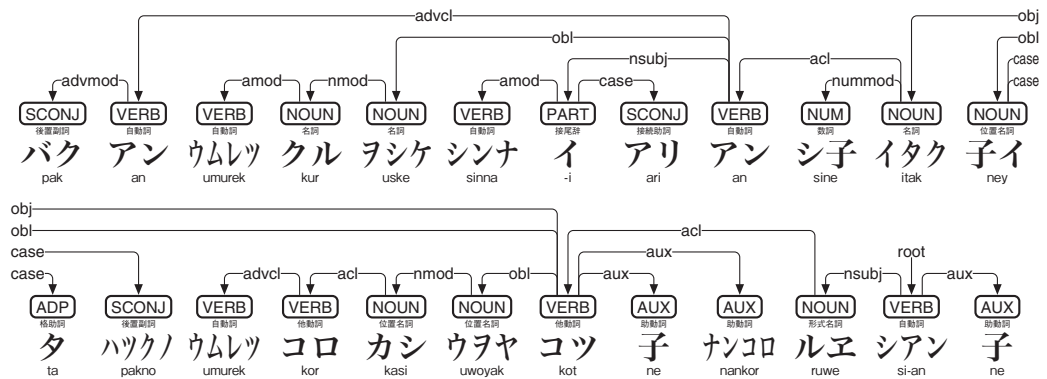
夫婦有別での「夫婦」のアイヌ語訳は、「ウムレツクル」で統一されている(3-1・3-3・3-6・3-9・3-12)が、3-13後半では「ウムレツクル」が「ウムレツコロ」に書き直されている。「夫」のアイヌ語訳は「ホク」で、「婦」のアイヌ語訳は「マツ」で、いずれも統一されている(3-7・3-8)。「別」のアイヌ語訳は「シンナイ」で統一されている(3-12・3-13)。「陰」のアイヌ語訳は、「チウブカモイ」(3-1・3-5)、「メウンチウブ」(3-5)、「メウンヲカケ」(3-2)、「メイ」(3-5)と多彩である。「陽」のアイヌ語訳も、「レクトコトカモイ」(3-1)、「レクトコカモイ」(3-5)、「ホブケチウブ」(3-5)、「ホブケ」(3-5)、「シリボフケ」(3-2)と多彩である。

3-3「セベンケ」には *se penke* を当てたが、これは今でも自信がない。3-6「ウヲヤコテ」には *uwoyak kote* を、3-13「ウヲヤコツ」には *uwoyak kot* を当てたが、やはり自信がない^[17]。3-10「ウエシカヒ」には *u-esikari* を当てたが、そうすると3-7「ウエシカビ」に *u-esikari pe* を当ててしまっていていいのか、疑義^[19]が残る。3-10・3-11「ツセ」には、悩みつつも *cise* を当てた。3-10・3-11「ヲソン」には *osoyun* を当てたが、疑義^[15]が残る。

アイヌ語訳『五倫名義解』の父子有親・君臣有義・夫婦有別を、ざっと通して見ると、各章末のキメ言葉にあたる「掟たるべし」が、父子有親1-12では「レンカエ子アネクニタバンナ」、君臣有義2-14では「カシア子ラツチケクンベタバンナ」、夫婦有別3-13では「カシウヲヤコツ子ナンコロエシアン子」と統一感が無く^[20]、キメ言葉になっていない。一方、父子有親1-10では「カシア子ラツチケクンベタバン」を「本意とすべし」のアイヌ語訳としており、近代日本語とアイヌ語の間で多対多^[4]となっている。

[19] 浅井亨: 加賀屋文書の中のチャコルベ, 北方文化研究, 第6号(1973年3月), pp.131-162.

[20] キメ言葉の「掟たるべし」は、長幼有序では「ウヲヤトモコテ子イルエシアン子」、朋友有信では「カシウヲヤトモコテ子ナンコロエタハンナ」であり、*uwoya tumu kote* へ収斂しているようにも見える。



その直前のキメ言葉「不易の法と定め給ふ」は、父子有親 1-11 では「モシリカシタエヨマレマツ子シトモヌヒタバナ」、君臣有義 2-13 では「ムシリカシケヲマイタカシカモイ子ウヲヤコテアンルエ子ナ」、夫婦有別 3-12 では「子イバツクノイタツカシソモヲビツタカムイ子アヌウコアンテアンルエタハン」と、どんどん長くなって^[21]いる。

もう一つのキメ言葉「倫を定るなり」は、父子有親 1-3 では「ウカトコロールエタバン」、君臣有義 2-2 では「ウカトコロールエタバン」、夫婦有別 3-6 では「ウカトウヲヤコテルエタバン」である。ただ、1-3・2-2「ウカトコロ」に u-katkor を当てるなら、3-6「ウカトウヲヤコテ」に o katu uwoyak kote を当てていいのか、かなり悩ましい^[22]ところである。『五倫名義解』のテーマを「倫」とするなら、これらの部分では、「ウカト」という表現で「倫」のアイヌ語訳に挑戦している、とも見える。

一方で、1-1 の「五倫」^[23]は、アイヌ語訳が「五倫アリバアセエラマツエツケウエ」となっている。「倫」と「五倫」で表現が異なっているのだが、これを検討するためには、アイヌ語訳『五倫名義解』の残り部分(長幼有序・朋友有信・空谷茂潤による刊記)も作業を進める必要がある。そう考えて、次の長幼有序に着手したが、残念ながら本稿のべ切に間に合わなかった。われわれの当日の発表に期待されたい。

^[21]「不易の法と定めたまふ」は、長幼有序では「マウエシヨモベ子クニエタクマコツアンルエ子」、朋友有信では「子イバツクノソモベ子クニイタクマコツアンルエタバン」であり、somo péne kuni itakmakkusute=an へ収斂しているようにも見える。

^[22]類似の表現として「倫を立置て」があり、長幼有序では「ウカトア子ウコアンテ」、朋友有信では「ウカトアノコアンテワ」である。「ウカト」に o katu を当てても大丈夫そうだが、なお検討を要する。

^[23]これ以外に「五倫」が2つ、刊記に現れるが、1つは「アシキ子エキン子ウエメキテ」と丸囲みで訳されており、もう1つは空欄となっている。しかし、1-1 では「アシキ子エキン子ウエメキワ」を「分て五つにして」のアイヌ語訳としており、「五倫」の訳とはしていない。

付録 Universal Dependencies の概要

UD は、書写言語における品詞・形態素属性・依存構造(係り受け関係)を、言語に関わらず記述する手法^[24]である。句構造を考慮せずに係り受け関係を記述することで、言語横断性を高めており、全ての文法構造を単語間のリンクで記述するのが特徴である。

依存構造解析それ自体は、Tesnière の構造的統語論^[25]に源を発し、Мельчук の有向グラフ記述^[26]によって、一応の完成を見た手法である。その最大の特長は、いわゆる動詞中心主義によって言語横断的な記述が可能だという点にあり、Мельчук 依存文法をコンピュータ向けに洗練した UD においても、言語に関わらない記述、という特長が前面に押し出されている。UD における文法構造記述は、句構造を考慮せず、全てを単語間のリンクとして表現する。これにより、言語横断的な文法構造記述を可能としている。

UD 係り受けコーパスの交換用フォーマットとして、CoNLL-U と呼ばれるタブ区切りテキスト(文字コードは UTF-8^[27])が規定されている。CoNLL-U の各行は各単語に対応しており、表 1 に示す 10 個のタブ区切りフィールドで構成される。ID・FORM・LEMMA は、単語そのものに関するフィールドである。UPOS・XPOS・FEATS は、単語の品詞と形態素属性に関するフィールドである。HEAD・DEPREL・DEPS は、単語の係り受けに関するフィールドである。

UD における係り受け関係は、単語間の有向グラフを HEAD と DEPREL で記述する。HEAD は、その単語に入る有向枝のリンク元 ID を示しており、DEPREL は、その有向枝における係り受けタグである。ただし、HEAD が 0 の場合、その枝に入るリンク元は存在しない。リンクの本数は単語の個数に等しく、各リンクのリンク先は、全て互いに異なっている。すなわち、各単語から出るリンクは複数の可能性があるが、各単語に入るリンクは 1 つだけである。なお、リンクはループしない。

UD の係り受けリンクは、Мельчук 依存文法の後裔にあたり、いわゆる動詞中心主義である。動詞をリンク元として、主語や目的語へとリンクする。修飾関係においては、被修飾語から修飾語へとリンクする。ただし、側置詞(前置詞や後置詞)を体言の修飾語だとみなす点^[28]が、Мельчук とは異なっている。また、コンピュータ文においては動詞中心主義を採らず、補語をリンク元として、主語や繫辞へとリンクする。

なお、UD は単語長を規定しておらず、各言語ごとに、自由に単語長を決めることができる。本稿の近代日本語 UD では、国語研短単位^[29]を単語長として用いた。また、アイヌ語 UD では『アイヌ語沙流方言辞典』^[30]を、作業上の単語認定に用いた^[9]。

^[24]Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman: Universal Dependencies, Computational Linguistics, Vol.47, No.2 (June 2021), pp.255-308.

^[25]Lucien Tesnière: Éléments de Syntaxe Structurale, Paris: C. Klincksieck (1959).

^[26]Igor A. Mel'čuk: Dependency Syntax: Theory and Practice, New York: State University of New York Press (1988).

^[27]ISO/IEC 10646-1:1993/Amd.2:1996 Information Technology — Universal Multiple-Octet Coded Character Set (UCS) — Part 1: Architecture and Basic Multilingual Plane, Amendment 2: UCS Transformation Format 8 (UTF-8), International Organization for Standardization, Genève (October 15, 1996).

^[28]Joakim Nivre: Towards a Universal Grammar for Natural Language Processing, CICLing 2015: 16th International Conference on Intelligent Text Processing and Computational Linguistics (April 2015), pp.3-16.

^[29]近藤明日子: 近代文語 UniDic 短単位規程集, Ver.1.1, 立川: 国立国語研究所コーパス開発センター (2016 年 3 月).

^[30]田村すず子: アイヌ語沙流方言辞典, 東京: 草風館 (1996 年 9 月).

表 1: CoNLL-U の各フィールド

1. ID: 単語ごとに付与されたインデックスで、文ごとに 1 から始まる整数。縮約語に対しては、単語の範囲を示すの也可。
2. FORM: 語、または、句読記号。
3. LEMMA: 基底形、語幹。
4. UPOS: UD で規定された言語普遍的品詞タグ (表 2)。
5. XPOS: 言語固有の品詞タグ。
6. FEATS: UD で規定された言語普遍的形態素属性のリスト。言語固有の拡張も可。
7. HEAD: 当該の単語の係り受け元 ID。係り受け元が無い場合は 0 とする。
8. DEPREL: UD で規定された言語普遍的係り受けタグ (表 3)。HEAD が 0 の場合は root とする。言語固有の拡張も可。
9. DEPS: 複数の係り受け元を持つ場合、全ての HEAD:DEPREL ペア。
10. MISC: その他のアノテーション。

表 2: UD 品詞タグ (UPOS)

Open class words	Closed class words	Other
ADJ 形容詞	ADP 側置詞	PUNCT 句読点
ADV 副詞	AUX 助動詞	SYM 記号
INTJ 感嘆詞	CCONJ 並列接続詞	X その他
NOUN 名詞	DET 限定詞	
PROPN 固有名詞	NUM 数詞	
VERB 動詞	PART 接辞	
	PRON 代名詞	
	SCONJ 従属接続詞	

表 3: UD 係り受けタグ (DEPREL)

	Nominals	Clauses	Modifier Words	Function Words
Core arguments	nsubj 主語 obj 目的語 iobj 間接目的語	csubj 節主語 ccomp 節目的語 xcomp 節補語		
Non-core dependents	obl 斜格補語 vocative 呼称語 expl 形式語 dislocated 外置語	advcl 連用修飾節	advmod 連用修飾語 discourse 談話要素	aux 動詞補助成分 cop 繫辞 mark 標識
Nominal dependents	nmod 体言による連体修飾語 appos 同格 nummod 数量による修飾語	acl 連体修飾節	amod 用言による連体修飾語	det 決定語 clf 類別語 case 格表示
Coordination	MWE	Loose	Special	Other
conj 接続 cc 接続語	fixed 固着 flat 並列 compound 複合	list 細目 parataxis 隣接表現	orphan 親なし goeswith 泣き別れ reparandum 言い損じ	punct 句読点 root 親 dep 未定義