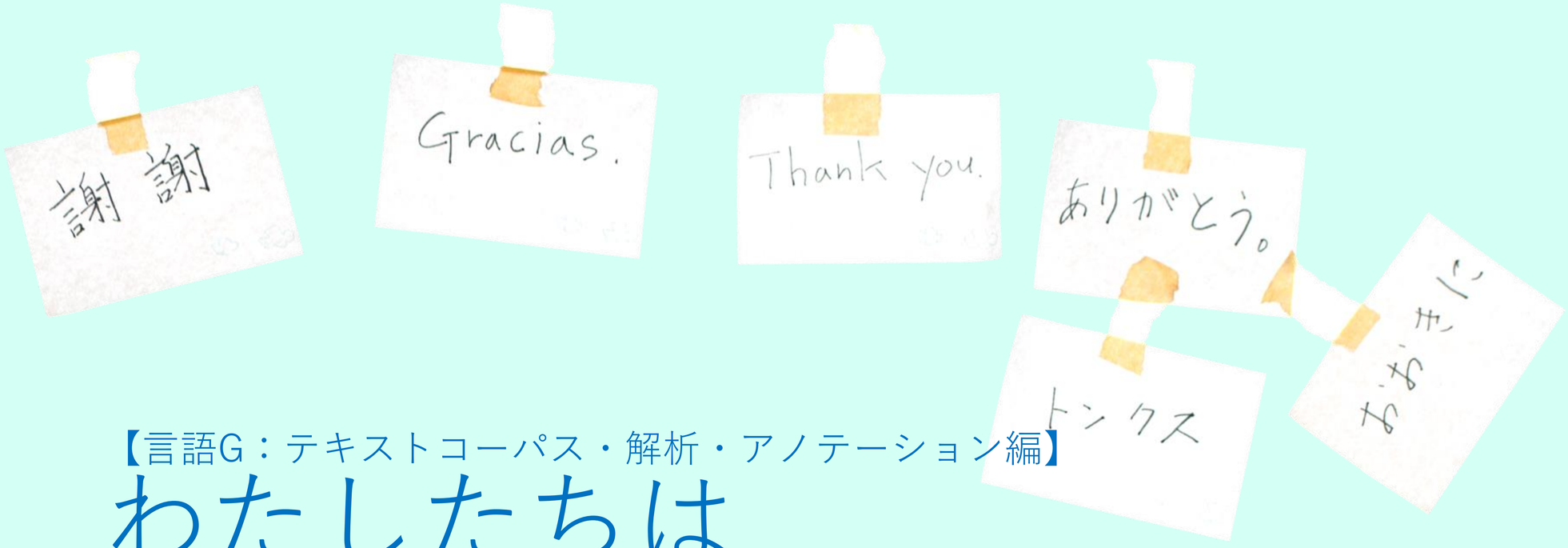
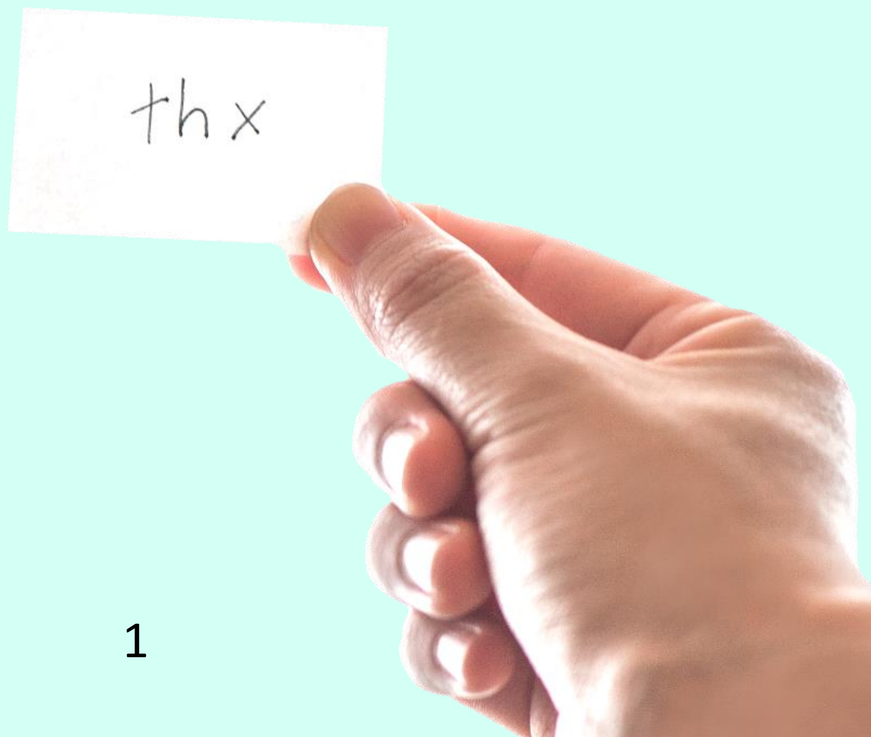


【言語G：テキストコーパス・解析・アノテーション編】

わたしたちは プロダクトは持っていませんが、 データ収集・解析の豊富な ノウハウを持っています



• テキストデータの作成と解析

　現在インターネット上には、無数の文が公開されています。
　しかし、自分が本当に望む文というのは、僅かしかない場合や、量はあっても、そのままでは使えないということが多くあります。
　文を収集または作成するだけではなく、
　お客様が本当に必要な情報は何か？を
経験豊富なアイアール・アルト社員が一緒に考えていきます。

• 音声データの収録と解析

　アイアール・アルトでは、お客様が望む形の音声を収録するだけでなく、
　音声データの解析・分析を行います。発話内容の書き起こしはもちろん、分節音／音素のセグメンテーションなど、音響情報をもとに、専門家が行う分析を代行することができます。
日本語だけでなく、外国語の分析も可能です。

• 映像データの収録と解析

　映像（動画）データには多くの情報が含まれていますが、そのままでは扱いにくいものです。
　例えば、被写体の人物が何を見ているか（視線）ということや、
何をしているのか（動作）については、機械が認識しやすい形で別途情報を付与しなければ、学習用データとして用いることは困難です。
　アイアール・アルトでは、映像収録から解析結果の情報付与まで、一貫して行います。

アイアールアルトの作業プロセス

お客さま

要件定義

データ作成作業

- 仕様を理解したプロジェクト担当者がプロジェクトを担当
- 適切な作業者を探す
- 仕様を作業可能、かつ、揺れが出にくいようなガイドラインに落とし込む
- 作業内容と品質の管理
- 納品まで責任を持って進捗管理

データのチェック

- 必要な数量を満たしているか
- 意図したデータになっているか
- 不要なノイズはないか

納品

アルトメンバー: 約5000名の作業者プール

見積相談、お気軽に

複雑な工程の設計も、経験豊富なアイアール・アルトなら、力になります。

お見積もり提示の段階で、作業手順を含めて、ご提案をいたします。

例：　①手順書のブラッシュアップ、作業ツールの提案
②発話を分析単位に整形（機械処理）
③ラベル付与（アノテーターによる判定）→④中間納品
④チェック作業

データの品質を決めるチェック作業は、データベース上で行う、機械的なチェックはもちろんのこと、人手で行うチェックをかならず工程に含めます。
チェック作業をどの段階に入れるか、誰が、どのように行うか、施業手順の設計段階で決定します。

例：

- **クロスチェック**
- **ダブルチェック**
- **異なり作業**
- **抜き取りチェック**
- **（外国語の場合）ネイティブによるチェック　　等**

実績のご紹介

言語学の専門知識が必要なコーパスも、アイアール・アルトならお手伝いできます。

① 形態素解析
② 音声分析
③ 修辞構造木
④ 対話データアノテーション

①形態素解析

固有表現や専門用語や砕けた表現などの**形態素解析の誤り**について
お困りの際にはご相談ください

現状	提案	実現
形態素解析誤りが原因で何らかの処理に困っている	形態素解析結果を加工します	分析や開発が可能に！
機械学習用の正解データがない	**対象分野の品詞タグ付コーパス**を構築します	**機械学習**が可能に！
分析したい語が**未知語**である	**未知語の形態素辞書**を分析したい言語データから作ります	**未知語**も分析が可能に！
異表記があるため正確な統計分析が難しい	異表記には**代表表記**を付与します	**表記の違い**を気にせず正確な分析が可能に！

形態素解析の単語分割の事例

侍ジャパンの村上ちゃんがWBCでホムランキメたのぉ～！
村神様マジ神ってるぅ～！

	解析器(&辞書)	単位区切りの例
mecab	mecab (ipa)	| 侍 | ジャパン | の | 村上 | ちゃん | が | WBC | で | ホムランキメ | た | の | ぉ | ～！ | | 村 | 神様 | マジ | 神 | っ | てる | ぅ | ～！ |
	mecab (UniDic Short)	| 侍 | ジャパン | の | 村上 | ちゃん | が | WBC | で | ホムランキメ | た | のぉ | ～ | ！ | | 村 | 神 | 様 | マジ | 神 | っ | てる | ぅ | ～ | ！ |
	mecab (UniDic Long)	| 侍 | ジャパン | の | 村上ちゃん | が | WBC | で | ホムランキメ | た | のぉ | ～ | ！ | | 村神様 | マジ | 神 | っ | てる | ぅ | ～ | ！ |
	mecab (neologd)	| 侍ジャパン | の | 村上 | ちゃん | が | WBC | で | ホムランキメ | た | の | ぉ | ～！ | | 村神 | 様 | マジ | 神 | って | るぅ | ～！ |
sudachi	Sudachi (C単位)	| 侍 | ジャパン | の | 村上 | ちゃん | が | WBC | で | ホムランキメ | た | のぉ | ～ | ！ | | 村神 | 様 | マジ | 神 | っ | てるぅ | ～ | ！ |
juman	juman++	| 侍 | ジャパン | の | 村上 | ちゃん | が | WBC | で | ホムラン | キメた | のぉ～ | ！ | | 村 | 神様 | マジ | 神 | って | るぅ～ | ！ |
編集例	短い単位	| 侍 | ジャパン | の | 村上 | ちゃん | が | WBC | で | ホム | ラン | キメ | た | のぉ～ | ！ | | 村神 | 様 | マジ | 神っ | てるぅ～ | ！ |
	長い単位	| 侍ジャパン | の | 村上ちゃん | が | WBC | で | ホムラン | キメ | た | のぉ～ | ！ | | 村神様 | マジ | 神っ | てるぅ～ | ！ |

短めの形態素解析結果の加工事例

侍ジャパンの村上ちゃんがWBCでホムランキメたのぉ～！

表層形態素	語彙素	品詞	活用型	活用形	固有表現
侍	侍	名詞-普通名詞-一般			B-ORGANIZATION
ジャパン	ジャパン-Japan	名詞-固有名詞-地名-国			I-ORGANIZATION
の	の	助詞-格助詞			
村上	ムラカミ	名詞-固有名詞-人名-姓			B-PERSON
ちゃん	ちゃん	接尾辞-名詞的-一般			
が	が	助詞-格助詞			
WBC	ＷＢＣ	名詞-普通名詞-一般			B-EVENT
で	で	助詞-格助詞			
ホム	ホーム-home	名詞-普通名詞-一般			
ラン	ラン	名詞-普通名詞-サ変可能			
キメ	決める	動詞-一般	下一段-マ行	連用形-一般	
た	だ	助動詞	助動詞-ダ	終止形-一般	
のぉ～	のー	助詞-終助詞			
！	！	補助記号-句点			

目的に応じて適切な**形態素単位**に加工します
（**品詞**や**固有表現ラベル**の付与や異表記に**代表表記**を付与することも可能です）

長めの形態素解析結果の加工事例

村神様マジ神ってるぅ～！

表層形態素	語彙素	品詞	活用型	活用形	固有表現
村神様	**村上宗隆**	名詞-固有名詞-人名-一般			B-PERSON
マジ	まじ	形状詞-一般			
神っ	神る	動詞-一般	五段-ラ行	連用形-促音便	
てるぅ～	てる	助動詞	下一段-タ行	終止形-一般	
！	！	補助記号-句点			

語彙素のアレンジも可能です
（「村神様」は東京ヤクルトスワローズ村上宗隆選手の愛称）

「神ってる」など**新語**も文法的に**正しい品詞情報**の編集も可能です
（「神ってる」とは「神がかっている」（神懸っている）という意味）

音声分析もおまかせください

収録された音声に分析（ラベリング）を加えることで、音声合成・音声認識エンジンを向上させます。

・**分節音／音素ラベリング**

音響情報をもとに分節音／音素のセグメンテーションを行う。

過去の実績：日本語／英語／中国語（声音・韻母単位）

・**韻律ラベリング**

聴覚情報または音響情報をもとに、韻律のラベリングを行う。

過去の実績：日本語のアクセントおよびイントネーション

英語の強勢

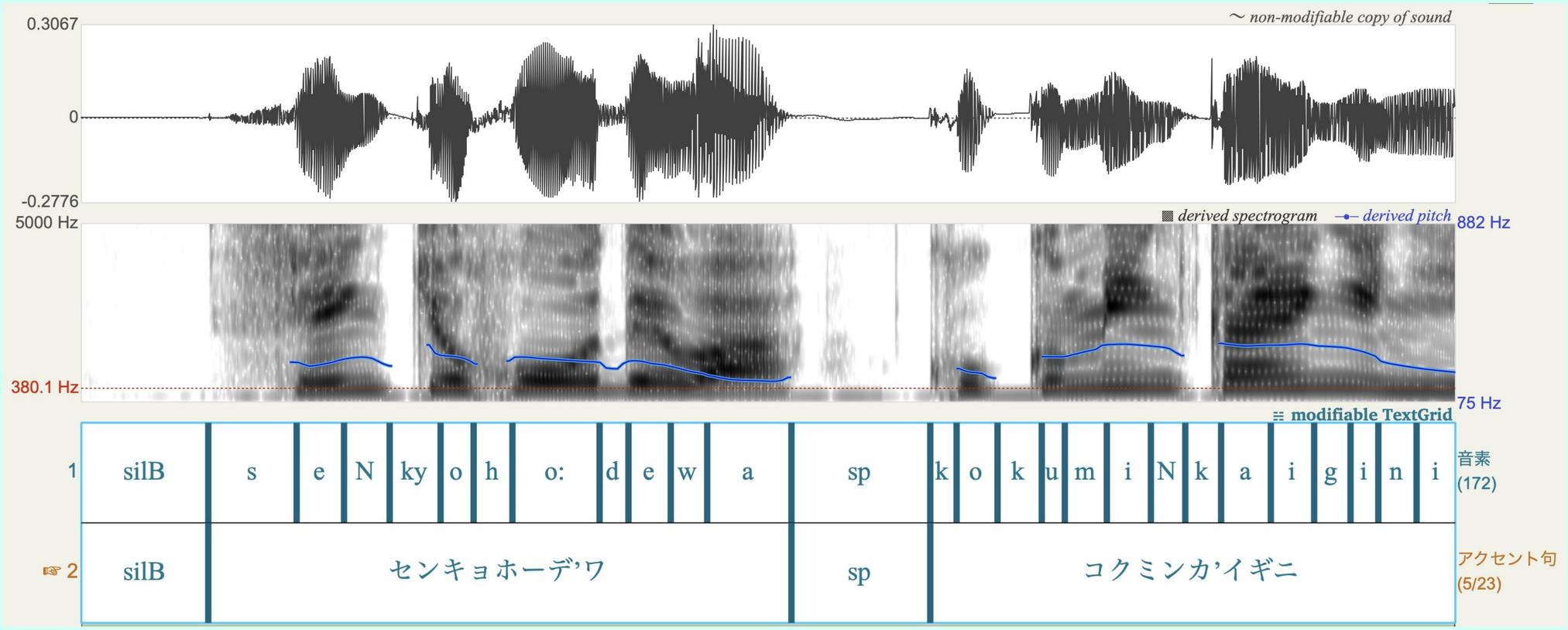
・**パラ言語等のラベリング**

分節音や韻律のラベリングと同時に、パラ言語となりうる、音響イベントのラベリングを行う。

過去の実績：笑い声／息継ぎ／ポーズ

音声分析もおまかせください

- 日本語の音声に対し、音素情報・アクセント情報・ポーズ情報を付与した例
  ※Praat（音声：JSUT v. 1.1 BASIC5000）



音素ティアー

音素・無音区間のラベルは音声認識エンジン「Julius」と同様

休止の直後に無声子音が続く際、クロージャーは休止区間に含めた

アクセント句ティアー

カタカナの発音綴りで表記

1区間に1つのアクセント句

アクセント核は「’」で表した

修辞構造木の構築事例

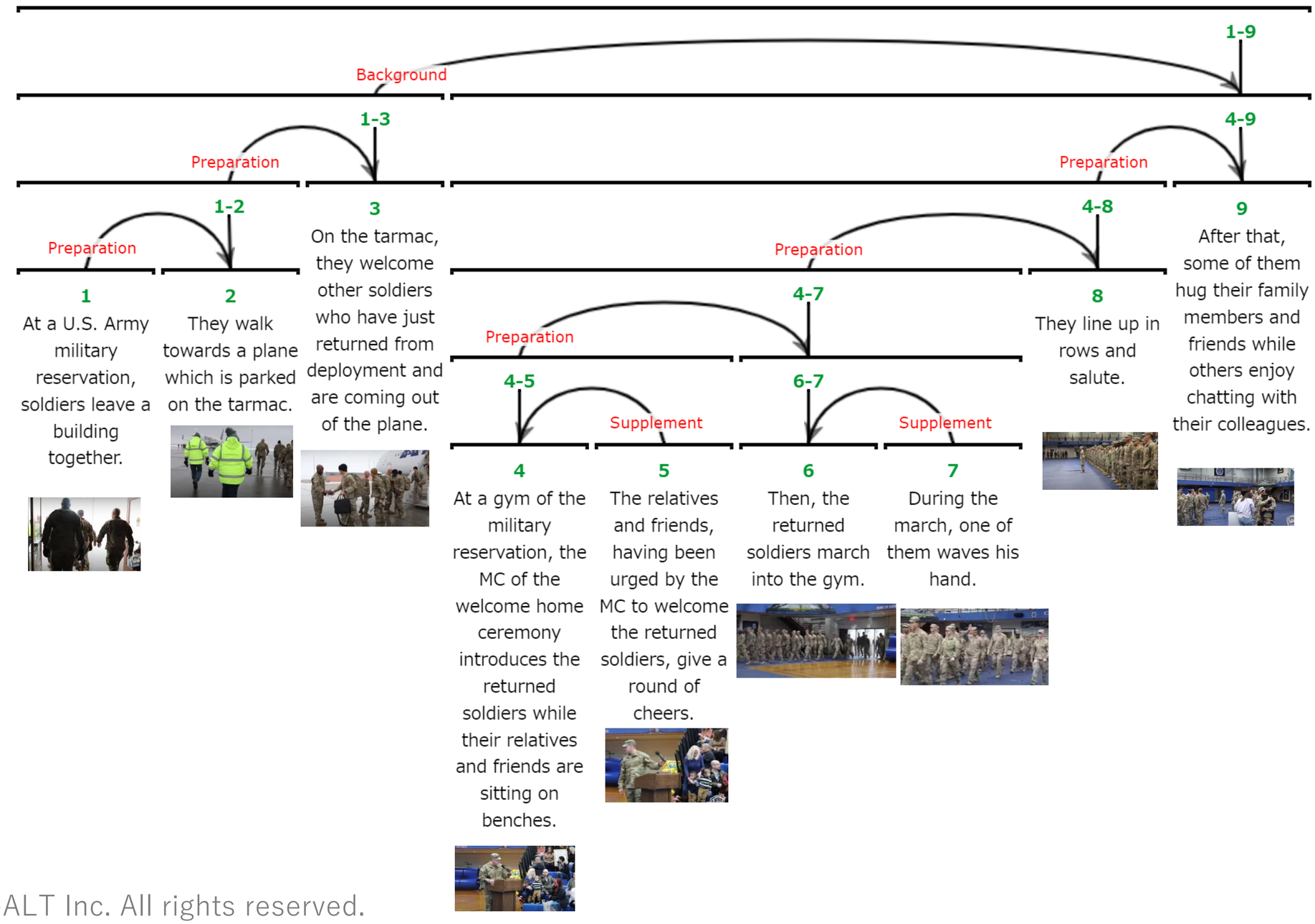
動画中で何が起こっているのか、テキスト化する（キャプション文）だけでなく、イベント間の関係を構造化します。

- 工程①：1-2 分程度の 動画を選定
- 工程②：動画で「何が起きているか」を記述した数文の英文作成
- 工程③：動画上で、各英文で記述されたイベントが発生する区間の時間を取得
- 工程④：英文全体に対して
  修辞構造木のアノテーションを付与

作業工程は、見積もり提案時にじっくりと相談・検討…。

修辞構造木の構築事例

任務から帰還した兵士たちが基地で家族や友人と再会を果たす動画
　イベント間の関係を、修辞構造（因果関係など）で表すデータを人手で作成しています。



弊社がお手伝いしたコーパスをご紹介します。

修辞構造木の構築に関連するもの

- 科学研究費助成事業データベース
「動画談話構造解析とそれを用いた要約生成」

- 言語処理学会（2021年3月）
「動画の談話構造解析」

対話データアノテーションの構築事例

案件の目的：対話における各発話の修辞機能を分析

ロボットとの自然な「雑談」のためには、人間が意識せずに行っている、言語行為のモデル化が必要

修辞機能とは：話し手書き手が発信する際に、 言及する対象である事態や事物、人物等を捉え表現する様態を分類し概念化したもの

- 工程①：手順書のブラッシュアップ
- 工程②：発話を分析単位（節）に整形
- 工程③：各節に対して６種類以上のラベルを付与
- 工程④：チェック・品質管理

対話データアノテーション

発話を分析単位（節）に整形

でもね、
小さくなるから便利ですよ。

でもね、小さくなるから
便利ですよ。

分析単位が「節」であれば
節単位に分割・結合

対話データアノテーション

各文に対して6種類以上のラベル・注釈を付与

topic	speaker	節	文タイプ	下位分類	時間要素	空間要素
レジ袋	A	小さくすると	従属	条件節		
レジ袋	A	便利ですよ。	主節		習慣・恒久	状況外
レジ袋	A	女房に教わりました。	主節		過去	参加
レジ袋	B	丸めてる、私も。	主節		習慣・恒久	参加
レジ袋	C	今度やってみようっと。	主節		未来意志	参加

弊社がお手伝いした
コーパスをご紹介します。

展示許可を得たコーパスについて、掲載しております。

title	概略	規模	言語	アイアール・アルトの役割	研究利用について
高齢者の語りのコーパス	●書き言葉だけでなく、音声発話とその書き起こしも収録し、さらに、認知症テスト（長谷川式知能評価スケール）の結果も合わせて収載しています。	３０名の高齢者（平均年齢７８歳：認知症予備群７名含む）の音声発話（平均２０分）とその書き起こし作文（平均５００文字）から構成されます。	Japanese	●高齢者アサイン ●弊社収録室での実施 ●データ整理	国立大学法人奈良先端科学技術大学院大学ソーシャルコンピューティング研究室（荒牧英治准教授）までお問合せください（http://sociocom.jp/）
MedNLPDoc電子カルテコーパス	●日本語電子カルテ文章を用いたシェアードタスク「MedNLPDoc」にて配布された電子カルテ文章です。 ●日本語のテキスト診断データに、適切な診断名とICDコードが付与されています。 ●医療現場におけるカルテの電子化に伴い、大規模な医療情報の利活用が期待されています。 データ記録形式の標準化が必須となり、それを効率よく行うための言語処理が注目されています。 ●複数の病名コードを持ちうる診療データを扱うこのタスクは、文章に対するマルチ・ラベリング問題と位置付けられます。	80 documents	Japanese	●医療専門家のコーディネート ●アノテーション監督	NTCIR事務局にお問い合わせください（http://research.nii.ac.jp/ntcir/index-en.html）
MedWebツイッタータスクコーパス	●8つの病気または症状（インフルエンザ、花粉症 等）に関するツイートデータです。 ●Twitterから収集したツイートデータの再配布は禁止されているため、クラウドソーシングにより作成しました。 ●英語と中国語のコーパスは、日本語で作成した模擬ツイートデータを翻訳して構築しています。	学習データ1,920 発言 テストデータ640 発言 （計2,560 発言）	Japanese English Chinese	●模擬ツイートデータ作成 ●翻訳	NTCIR事務局にお問い合わせください（http://research.nii.ac.jp/ntcir/index-en.html）

国立情報学研究所ではNTCIR（エンティサイル、NII Testbeds and Community for Information access Research)プロジェクトという活動を進めています。NTCIRでは、人工知能の発達のために、有効性の検証とベンチマークのためのデータセット構築を行っています。私たちは2010年から成果報告会のスポンサーとして参加してきました。公開されているデータのうちいくつかについては、実際のコーパス作成に関わらせていただきました。