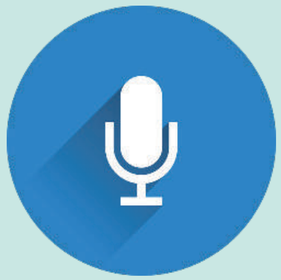
わたしたちは
プロダクトは持っていませんが、
データ収集・解析の豊富な
ノウハウを持っています

• 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。

• 音声データの収集

　テキストに比べ、音声は個別の用途に合ったものがなかなか存在しません。
　音声は年齢、性別、居住地の他、収録環境によってもデータの性質が大きく変わってしまいます。
　アイアール・アルトでは、お客様が望む形の音声を収録致します。
　お客様の希望に合わせて、海外での収録も行えます。



• 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。

こんな時ご相談ください

■既存のクラウドソーシングの利点：

- １）一度に　２）低予算で　３）地理的制限なしに　４）大量のデータを集めることが可能

■しかし既存のクラウドソーシングが不向きなケースも：

- データ提供者の属性を絞る場合（幼児や高齢者、特定言語の話者など）
- 作業者に対し高度な訓練が必要な場合
- 選別・加工されたデータが必要な場合
- データの量よりも質が重要な場合
- 一貫性の必要なデータを長期にわたって収集する場合

→　ご相談ください

IR-ALTを選ぶ理由

■あらゆるデータに対応

- 多種多様な属性のワーカーと独自ネットワークで繋がっています
- 複雑なタスクや長期の収集期間にも対応できます

■精度の高いデータ

- チェック作業を重視し、テクノロジーとノウハウを駆使して精度の高いデータを作ります

■様々な作業ツールで対応

- 紙からWebアプリまで、タスクに最適な作業ツールをご用意します

■安心のフォロー体制

- 決まった担当者がつき、詳細に要件をヒアリングするため、きめ細やかなフォローが可能です

作成データ例

タスク例：質問と答えで構成される英語対話の2往復目を作成してください	
Q1: Are you good at using a computer? A1: Yes, I am. Q2: ok ,,,how much time you spent on computer?? A2: about 8 to 9 hours.	Q1 : Are you good at using a computer? A1 : Yes, I am. Q2 : How much time do you spend staring at screens each day? A2 : About 8 to 9 hours.
Q1 : About how many New Year’s cards do you exchange? A1 : I wrote about 100. Q2 : son muchas tarjetas A2 : si pero siempre reparto asi	Q1 : About how many New Year’s cards do you exchange? A1 : I wrote about 100. Q2 : Isn't that quite a lot? A2 : I guess I just have a lot of friends.
左：既存のクラウドソーシング タスクに沿わないデータや誤字脱字も	**右：弊社作成データ 不適切なデータを減らすノウハウがあります**

タスク例：提示されたトピックを含む、続きを話したくなるような 日本語のつぶやきを作成してください	
トピック：食べ歩き 評価用文：食べ歩きのプロ	トピック：食べ歩き 評価用文：私は食べ歩きが趣味で、先週末も食べ歩いてきたんですよ。
トピック：Twitter 評価用文：twittreでまた問題行動をアップする人が現れました。	トピック：Twitter 評価用文：Twitterでまた問題行動をアップする人が現れました。

形態素解析結果の加工

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

『君の名は』は、２０１６年８月２６日公開の新海誠監督の長編アニメーション映画。

	解析器(&辞書)	単位切りの例
JUMAN	JUMAN	｜『｜君｜の｜名｜は｜』｜は｜、｜２０１６｜年｜８｜月｜２６｜日｜公開｜の ｜新｜海｜誠｜監督｜の｜長編｜アニメーション｜映画｜。｜
	JUMAN++	｜『｜君｜の｜名｜は｜』｜は｜、｜２０１６｜年｜８｜月｜２６｜日｜公開｜の ｜新海誠｜監督｜の｜長編｜アニメーション｜映画｜。｜
MeCab	MeCab (IPA)	｜『｜君｜の｜名｜は｜』｜は｜、｜２｜０｜１｜６｜年｜８月｜２｜６｜日｜公開｜の ｜新海｜誠｜監督｜の｜長編｜アニメーション｜映画｜。｜
	MeCab (UniDicShort)	｜『｜君｜の｜名｜は｜』｜は｜、｜２｜０｜１｜６｜年｜８｜月｜２｜６｜日｜公開｜の ｜新海｜誠｜監督｜の｜長編｜アニメーション｜映画｜。｜
	MeCab (UniDicLong)	｜『｜君｜の｜名｜は｜』｜は｜、｜２０１６年｜８月｜２６日公開｜の｜新海誠監督｜の ｜長編アニメーション映画｜。｜
	MeCab (neologd)	｜『｜君の名は｜』｜は｜、｜２０１６年｜８月｜２６日｜公開｜の｜新海誠監督｜の ｜長編アニメーション｜映画｜。｜

｜『｜君の名は｜』｜は｜、｜２０１６年８月２６日｜公開｜の

ARTIFACT　　DATE

｜新海誠｜監督｜の｜**長編アニメーション映画**｜。｜

PERSON　　名詞-一般

形態素解析結果の加工事例

『君の名は』は、２０１６年８月２６日公開の新海誠監督の長編アニメーション映画。

表層形態素	品詞大分類	品詞中分類	品詞小分類	品詞細分類	代表表記	固有表現タグ
『	記号	括弧開	*	*		
君の名は	名詞	固有名詞	一般	*	君の名は。	ARTIFACT
』	記号	括弧閉	*	*		
は	助詞	係助詞	*	*	は	
、	記号	読点	*	*	、	
２０１６年８月２６日	名詞	固有名詞	一般	*	２０１６年８月２６日	DATE
公開	名詞	サ変接続	*	*	公開	
の	助詞	連体化	*	*	の	
新海誠	名詞	固有名詞	人名	一般	新海誠	PERSON
監督	名詞	サ変接続	*	*	監督	
の	助詞	連体化	*	*	の	
長編アニメーション映画	名詞	一般	*	*	長編アニメーション映画	

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）

AI音声アシスタントを動かすのに必要なデータを作成します

Amazon EchoのAlexaやDocomoのSebastienなど…時代はAI音声アシスタントブーム。

「時間を教えて」と言っても、「時計持ってる？」と聞いても、どちらの尋ね方でも
アシスタントが時間を教えてくれるのは、アシスタントが発話の意図を正しく判定できているから。
しかし「AI」とはいえ、はじめから意図を理解できるわけではありません。
多くのアプリケーションでは、適切に整備したデータが必要とされています。
例えば、音声アシスタントは一般的に、「発話と意図」の組み合わせデータや、
スロットを埋めるワードの辞書等が必要です。

アイアール・アルトでは、データの収集、分類、整備をオーダーメイドでお手伝いしています。

発話	意図
こんにちは	あいさつ
やっほー	あいさつ
ハロー	あいさつ
{野菜名}を使ったレシピ教えて	レシピ検索
{野菜名}料理にはどんなのがある？	レシピ検索

意図	スロット名
あいさつ	なし
レシピ検索	野菜名

スロット名	野菜名	異表記・同義語
野菜名	人参	ニンジン, にんじん, キャロット
野菜名	じゃがいも	ジャガイモ, ポテト

チャットボットや対話システムの精度を高めるために必要なデータを作成します

「機械翻訳」　「音声合成」　「音声認識」　「AI（人工知能）」という技術・研究分野において、自然言語処理は大きな役割を担っています。

アイアール・アルトは、自然言語処理技術を支える「コーパス」の作成を得意としています。

チャットボットや対話システムが生み出すエージェントやキャラクタたちが
自然な対話を行うためには、精度の高いコーパスが欠かせません。
アイアール・アルトは、コーパス作成をはじめ、
自然言語処理の研究・開発に利用できる言語データの収集、作成、整備を得意としています。

音声、テキスト、動画
なんでも集めます
＋アノテーションします

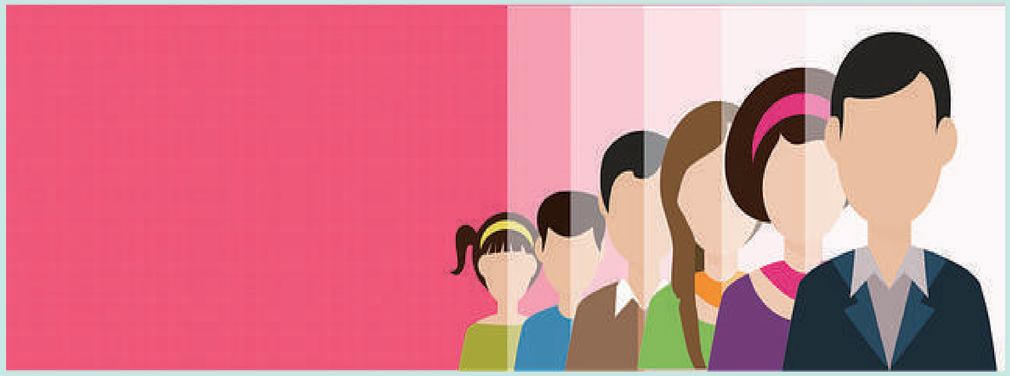
- やりたいことがあってデータは持っているけど、 前処理に手が回らない…
- やりたいことはあるけど、データを持っていない…

そんなお悩みをお持ちでしたら、
ぜひ一度お気軽にご相談ください。

医療など、専門分野のコーパスも作れます

コーパスのドメインによっては、専門家や限られた属性の方にしか作れないデータがあります。アイアール・アルトは、メールメンバー（協力希望者のML, 約5000名）や協力会社を通じて、難しいコーパス作成を可能とします。

メールメンバーは弊社のデータ作成に協力を希望する人々のメーリングリストです。シニア、子ども、外国語ネイティブなど、さまざまな属性の方が登録しています。



【実績例】
・音声認識のための子どもの声コーパス
・音声認識のための多言語発話コーパス
・雑談対話コーパス／質問コーパス

メールメンバーや協力会社を活用し、医療従事者、翻訳者、コールセンター勤務経験者など、コーパスドメインに合わせて、データ作成協力者をアサインすることができます。

【実績例】
・翻訳者による翻訳プロセステキストコーパス
・特定ドメインにおけるコールセンター音声コーパス
・日本語学習者による誤りテキストコーパス
・声優による感情音声コーパス

弊社がお手伝いした
コーパスをご紹介します。

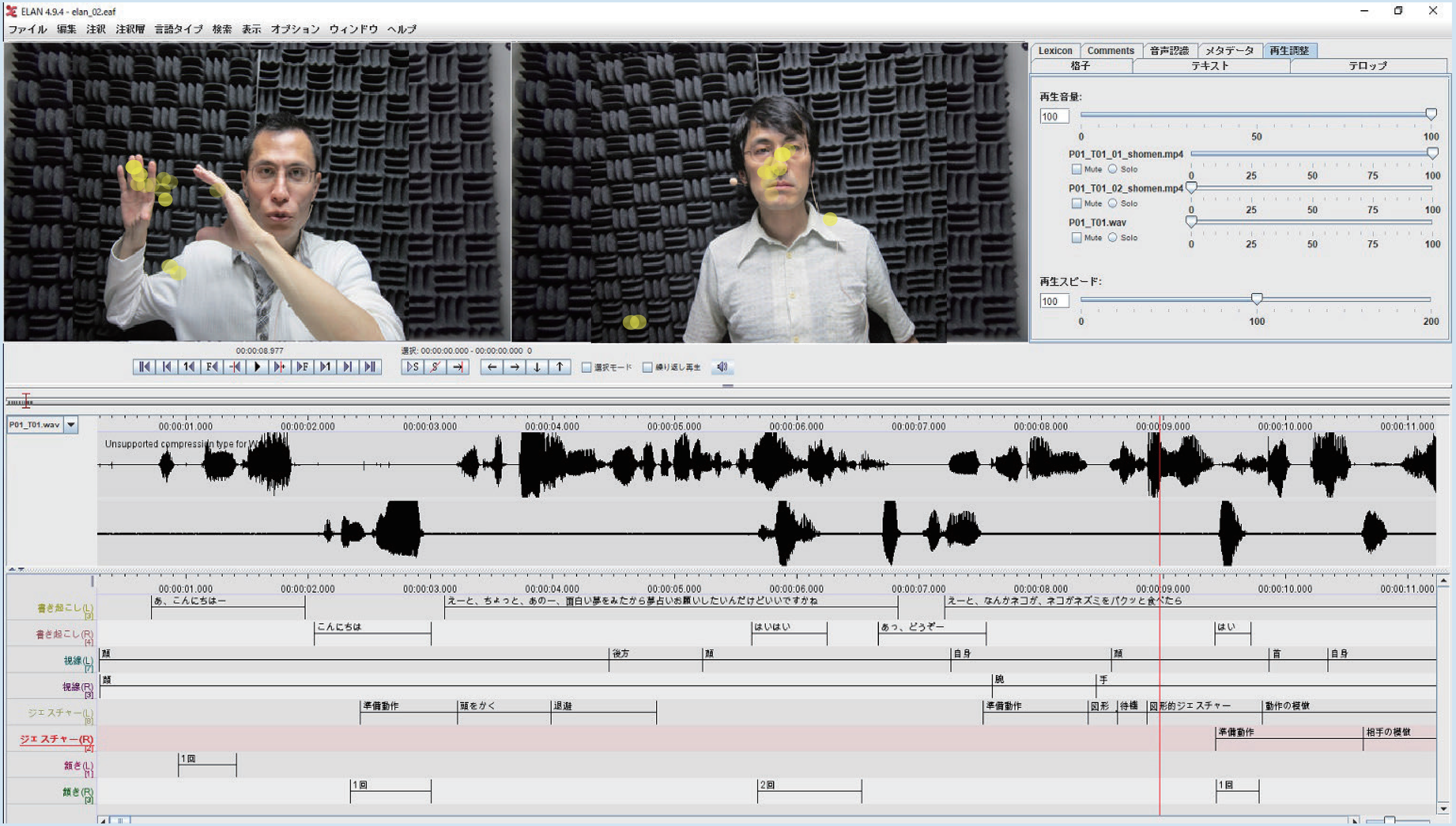
展示許可を得たコーパスについて、掲載しております。

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。

【マルチモーダルアノテーション例】

下記は、Elanを使用した発話・視線・ジェスチャーのアノテーションサンプルです。
写真上の黄色い点は、話し相手の視線の動きを可視化したものです。
※このデータは、展示会用に制作したサンプルです



弊社がお手伝いしたアプリをご紹介します。

展示許可を得たアプリについて、掲載しております。

【アプリ名】
ひとり英会話SiF
【リリース元】
株式会社サインウェーブ様
【アプリ概要】
最先端AIによる高い採点技術により、発音の評価が点数になるので、「話すチカラ」を高められます。また、インターネット環境があれば、24時間いつでもどこでも好きなタイミングで、自分のペースで学習を進めることができます。
【アイアール・アルトの役割】
アプリ内で使用される英語ネイティブによるお手本英語音声の収録。
(スクリプトの校正作業、英語ネイティブ話者のアサイン、
弊社収録室での収録実施、データ整理)
【アプリの利用】
App Store、Google Playからダウンロードしていただけます。
アプリ詳細は、QRコードをご参照ください。