言語資源小委(R8.6.30)
資　料　3

NINJAL
National Institute for Japanese Language and Linguistics

言語資源小委員会

話し言葉コーパスの現状について

小磯花絵（国語研究所）

2026年6月30日

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics

国語研究所における話し言葉コーパスの構築状況

CSJ 日本語話し言葉コーパス

独話(中心)

独話→ 会話

- NUCC 名大会話コーパス
- CWPC 職場談話コーパス
- CEJC 日本語日常会話コーパス

母語話者→学習者

- I-JAS 多言語母語の日本語学習者横断コーパス
- C-JAS 中国語・韓国語母語の日本語学習者縦断発話コーパス
- B-JAS 北京日本語学習者縦断コーパス

成人→子ども

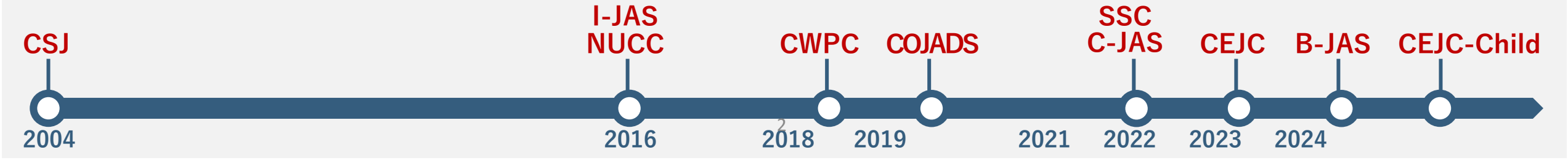
- CEJC-CHILD 子ども版日常会話コーパス

共通語→方言

- COJADS 日本語諸方言コーパス

平成・令和→昭和

- SSC 昭和話し言葉コーパス



公開方法1：オンライン検索システム「中納言」

コーパス名	略　称	個別検索	まとめて検索	備　考
書き言葉 現代日本語書き言葉均衡コーパス	BCCWJ	✅	✅	従来より利用いただいている BCCWJ のデータです（コーパスの紹介ページ）。こちらのページから BCCWJ アノテーションデータをダウンロードできます。
書き言葉 現代日本語書き言葉均衡コーパス 第2部	BCCWJ2	✅	✅	コーパスの紹介ページ
話し言葉 日本語話し言葉コーパス	CSJ	✅	✅	コーパスの紹介ページ
話し言葉 日本語日常会話コーパス	CEJC	✅	✅	コーパスの紹介ページ 有償版契約者は関連データを「データ配布」からダウンロードできます。
話し言葉 子ども版日本語日常会話コーパス	CEJC-Child	✅	✅	コーパスの紹介ページ
話し言葉 昭和話し言葉コーパス	SSC	✅	✅	コーパスの紹介ページ SSC の全データ（音声・転記・形態論情報・メタデータ等）をこちらからダウンロードできます。ダウンロードするには、コーパス追加利用の申請から昭和話し言葉コーパスの新しい規約に同意して利用を申請してください。
話し言葉 名大会話コーパス	NUCC	✅	✅	コーパスの紹介ページ
話し言葉 現日研・職場談話コーパス	CWPC	✅	✅	コーパスの紹介ページ
通　時 日本語歴史コーパス	CHJ	✅	✅	コーパスの紹介ページ
通　時 昭和・平成書き言葉コーパス	SHC	✅	✅	コーパスの紹介ページ
方　言 日本語諸方言コーパス	COJADS	✅	✅	コーパスの紹介ページ 関連データを「データ配布」からダウンロードできます。
日本語学習者 中国語・韓国語母語の日本語学習者縦断発話コーパス	C-JAS	✅	✅	コーパスの紹介ページ プレインテキストは「データ配布」からダウンロードできます。
日本語学習者 多言語母語の日本語学習者横断コーパス	I-JAS	✅	✅	コーパスの紹介ページ プレインテキスト・音声・作文は「データ配布１」からダウンロードできます。 I-JAS 外国語母語話者コーパス（I-JAS FOLAS）は「データ配布２」からダウンロードできます。
日本語学習者 北京日本語学習者縦断コーパス	B-JAS	✅	✅	コーパスの紹介ページ
オープン オープン CHJ	OpenCHJ	✅	✅	

検索結果の例

前後文脈最大300字
（学習者コーパスは100字）

※話者情報なし

<table>
<tr><th>会話ID</th><th>開始位置</th><th>連番</th><th>コア</th><th>前文脈</th><th>キー</th><th>後文脈</th><th>素読み</th><th>素</th><th>素細分類</th><th></th><th></th><th>型</th><th>形</th><th></th><th>ラベル</th><th>年齢</th><th>性別</th><th>出身地</th></tr>
<tr><td>T016_002</td><td>26410</td><td>15090</td><td>0</td><td>いく|ん|や # 外国|人|だ|ね # うん # いつ|も|ご飯|ヤン|どっち|か|出し|て|も # うーん # な # ん # まあ|外国|人|だ|から # 外国|人|だ|から # 思い切り|モンゴロイド|な|ん|だ|けど|ね # 顔|は|ね # うん # まあ|ね # 中国|の|富裕|層|みたい|な|顔|し|てる|けど # ん # よかっ|た|ね # 富裕|層|って|言わ|れ|て|ね # 痩せ|て|ない|もん|な # うーん # うん # うん # 優しく|ない|よ # いや # 痩せ|て|ない|から|な # んー # ど|し|た # 斎藤|さん # あー # うん # 何|言っ|てん|の # この|人 # んー # いや # なん|か|さ|髪型|が|斎藤|さん|ぽい|な|と|思っ|て # あー # そう|です|か # 一瞬|な # ありがと|ござい|ます # もう|ちょっと|ここ|剃れ|ば|さ|斎藤|さん|に|なれる|の|に|な # ありがとう # ありがとう|ござい|ます # 謙一|好き|だ|ね # その|ネタ # んー # なん|か|ね|面白く|ない|の|に|ね # あんた|ご飯|もう|ちょっと|いる # もう|いい # いら|ない|の # 食欲|あり|ませ|ん|ね # うそ|や # そう|っす|か # 誰|や|ねん # お前 # そう|な|の # お|野菜|食べ|ない|と # 便秘|便秘 # 便秘|に|なる|よ # ふーん # 便秘|っぴ # ない # 食っ|て|食っ|て # これ # で|も|味噌|汁|食べ|た|ね # 仕方ない|な # 仕方ない # 君|君|が|そんな|に|さ|そんな|に|食べ|て|ほしい|なら|食べ|て|あげる|よ # あの|あの|出汁|おいしく|なかっ|た # お|母|さん # おいしい # コ|あの|カツオ|の|やつ # そんな|に|さ|そんな|さ|ン # おいしい|おいしい # はい|はい # はい|はい # はい # あれ|なん|か|徳島|か|なん|か|の|じゃ|なかっ|た|っけ # そんな|に|食べ|て|ほしく|ない|ん|だっ|たら|食べ|なく|て|も|いい # んー # 味|が|結構|濃い|よ|ね # うん # ん # 塩味|も # うん # ん # うま # あの|トビウオ|の|出汁|と|か|売っ|て|た|けど|</td><td>ね</td><td># トビウオ|の|出汁|おいしい|ね # で|ね|すごい|高かっ|た|の # 高い|ね # ついで|に|俺っち|も|お|願い|し|やす # うん # なん|だろう|ね # あれ|すごい|高い|よ|ね # あれ|前|あ|あの|おいしい|トビウオ|の|出汁|どこ|で|買う|た|ん|やっ|た # ん # なん|で|オ|うち|ら|ウ|選ば|れ|た|ン|て # あれ|九州|や|っけ # あ # 違う # あれ|は|アゴ # え # なん|で|俺|ら|ウオ|選ば|れ|た|の # あ # アゴ|も|おいしかっ|た|な # うーん # ン|ふん # え # それ|って|別|に|トビウオ|も|なん|か|売っ|て|て|さ # 優秀|だ|から|じゃ|ない # うーん # で|も|アゴ|も|高い|よ|ね # うん # お|父|さん|以外|が # はい # そっ|す|ね # あと|あれ|も|売っ|て|た # ノドグロ # はい|はい # は|なん|か|ね|もう|ちょっと|ね|コ|こってり|って|ゆう|か # もっと|紳士|っぽく|食い|なさい # うん # 濃い|ん|だ|って # あんた|言っ|てる|意味|わから|へん # 味|が # ふーん # で|トビウオ|は|さっぱり|し|て|ん|だ|って # じゃあ|わたし|たち|は # たら|全部|食え|よ # 全部 # トビウオ|や|ね # トビウオ|か|な # で|も # わざわざ|ね|あーの|こう|ゆう|ふう|に|綺麗|に|食べ|て|あと|で|さ|また|綺麗|に|って|食う|って|ゆう|って|さ # 煮物|と|か|は|その|ノドグロ|が|合う|ん|だ|って|よ # ちょっと|コ|コク|が|ある|から # うーん # あー # うん # なるほど|ね # ん # ね # ちょ|自分|どう|ゆう|こと # だ|から|最初|に|ご飯|と|か|ね|卵|と|か|食っ|て|次|肉|と|こう|やっ|て|食べる|って|ゆう # うん # あんまり|そう|ゆう|の|食べ|た|こと|ない|から|わかん|ない # うん # で|も|アゴ|の|出汁|は|おいしかっ|た|よ|な # なん|や # 食べ|て|へん|やん # え # だ|から|だ|から|お|父|さん|に</td><td>ネ</td><td>ね</td><td></td><td>ネ</td><td>助詞-終助詞</td><td></td><td></td><td>雑談</td><td>IC02_理奈子</td><td>35-39歳</td><td>女性</td><td>東京都</td></tr>
<tr><td>T017_005a</td><td>92350</td><td>55510</td><td>0</td><td>一|回|は|吸っ|て|みる|でしょ # ほんと|に|二十歳|の|二三|年 # あ # みんな|そう|だ|よ|ね # うち|の|やつ|も|そう|だ # イ|一応 # 子供|が|できる|まで|は|吸っ|て|た|もん # うん # で|も|結婚|する|前|に|やめ|た|わ # え # あたし|なんか|子供|子供|いる|時|も|吸っ|て|た # あ # そう # 吸っ|て|た # あたし|やめ|た|の|は|何|年|だろう # 五十 # あ # そんな|ころ|吸っ|て|た # うん # そう|そ # 会社|で|ミン|みんな|一斉|に|やめ|出し|た|の|よ # それ|で|やめ|て|た # あ # そう # だっ|て|うち|ム|子供|も|ちょこっと|吸っ|て|た|の|よ # ふーん # あ # そい|で|そんな|に|吸っ|てる|わけ|じゃ|ない|けど # で|うち|の|壁紙|を|張り替え|て|そう|し|た|時|に|二人|で|さ|換気|扇|の|下|で|吸っ|て|て|さ # うん # で|ヤメ # そい|で|あたし|が|やめよ|か|な|っつっ|たら # あ # そんな|ころ|まで|吸っ|て|た|の # そう|そう|そう|そう # じゃ|コド|子供|が|おっきく|なる # うち|の|子|が|二十歳|すぎ|て|も|吸っ|て|た|から|イ # そんな # あ # そう|だっ|た|の # そんな|に|うち|の|子|は|好き|じゃ|なかっ|た|ん|だ|けど|ね # うん # やっぱし # クッ|ウ|付く # 周り # その|カー|壁紙 # すごかっ|た|よ # 壁紙|真っ黒|だっ|た # だ|から # 今|綺麗|だ|けど # こう|ゆう|とこ|は|禁煙 # いや # ここ|は|吸える|よ # あ # 吸える|ん|だ # うん # あー # あの|みんな|が|よけれ|ば|吸える|ん|だ|けど|も # 吸え|た # それ|か|エレベータ|の|脇|に # あ # 禁煙|の|は # あ # うん # 吸い|に|行く # キン|喫煙|室|が|ある # うん # うん|うん # あー # そう|な|ん|だ # うん # そう|そ # ここ|たばこ|吸え|ん|だ|よ # 今|大変|だ|よ|</td><td>ね</td><td># で|も|飲ん|でる|ヒ|飲ん|だら|吸い|たく|なる|ね # たばこ|吸っ|てる|人|は|ね # そう|だ|ね # うん # きっと|そう|だろう|ね # うーん # うん # うん|うん|うん|うん|うん # 禁煙|は|厳しい|よ|ね # たぶん|ね # うん # うん # ふーん # そう|だ|ね # 飲む|と|ね|吸い|たく|うん|なる|ね # 吸い|たく|なる|か|も|ね # そう|だ # 飲む|と|絶対|吸い|たく|なる # うん # なん|か|ね # うん # で|やめ|た|当初|は|で|も|なん|か|食べ|た|あと|が|吸い|たく|なっ|て|く|ん|の # うん # うん # 一服|が|なん|か|でき|ない|の|が # 食べ|た|あと|の|一服|は # え # え # 何 # え # 食べ|た|あと|の|一服 # え # 吸う|か|な # が # が|何 # ター|きつかっ|た|ね # でき|なく|やめ|た|時 # あー|あー|あー # やめ|た|時 # うん # そ|ゆう|時|どう|し|て|た|の # え # 騒い|で|た|の # あーっ|と # あー # や # 騒|が|ない # んー # どんどん|どんどん # ソー|そこ|まで|し|ませ|ん # あっ # ソン|そんな|あれ|は|し|ない|けど # お|肉|食べれ|ば # ムー # あたし|は|もう # お|肉|も|これ|も # ン|で|も|ヒ|ヒ|あの|なん|だっけ|あの|禁煙|パイポ|あれ|使っ|たら|逆|に|気持ち|悪い # 失礼|いたし|ます # そ|そ # だめ # だめ|だ|よ|ね # 禁煙|パイポ|と|ね # あれ|気持ち|悪い|よ|ね # うん # 何 # それ # その|なん|か|あの|ニコチン|を|ちょっと|取る|入れる|ん|だ|よ|ね # はい # こちら|北京|ダック|です # はい # その|こう|ゆう|たばこ|吸っ|た|気|に|なん|の # あれ|を|さ|もらっ|た|ん|だ|けど # あ # まだ|だ|よ # それ|は|まだ # あれ|は|だめ # はい # あたし|だめ|だ|と|思っ|た # はい # どれ # これ|取り|な # 食べ|な # あー # ある|ある # これ # レーレ # ん # ん # ん # ど|し</td><td>ネ</td><td>ね</td><td></td><td>ネ</td><td>助詞-終助詞</td><td></td><td></td><td>雑談</td><td>IC03_奥野</td><td>60-64歳</td><td>女性</td><td>東京都</td></tr>
</table>

一部のコーパス
音声再生可能
（CSJ, CEJC, CEJC-Child, SSC, COJADS）

※検索箇所のみ、ダウンロード不可

話者情報付き前後文脈も
表示可能だが範囲は狭い

詳細な文脈情報

T016_002	26110	14900	IC02_理奈子	よ	ヨ	よ		助詞-終助詞			ヨ	ヨ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26120	14910	IC02_理奈子	ね	ネ	ね		助詞-終助詞			ネ	ネ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26130	14920	IC01_慎吾	うん	ウン	うん		感動詞-一般			ウン	ウン	和	雑談	40-44歳	男性	大阪府	東京都
T016_002	26150	14930	IC01_慎吾	ん	ン	ん		感動詞-一般			ン	ン	和	雑談	40-44歳	男性	大阪府	東京都
T016_002	26160	14940	IC02_理奈子	塩味	シオアジ	塩味		名詞-普通名詞-一般			シオアジ	シオアジ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26180	14950	IC02_理奈子	も	モ	も		助詞-係助詞			モ	モ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26190	14960	IC04_ばあば	うん	ウン	うん		感動詞-一般			ウン	ウン	和	雑談	65-69歳	女性	奈良県	東京都
T016_002	26210	14970	IC01_慎吾	ん	ン	ん		感動詞-一般			ン	ン	和	雑談	40-44歳	男性	大阪府	東京都
		14980	IC01_慎吾	うま	ウマイ	旨い		形容詞-一般	形容詞	語幹-一般	ウマ	ウマ	和	雑談	40-44歳	男性	大阪府	東京都
		14990	IC02_理奈子	あの	アノ	彼の		連体詞			アノ	アノ	和	雑談	35-39歳	女性	東京都	東京都
		15000	IC02_理奈子	トビウオ	トビウオ	飛び魚		名詞-普通名詞-一般			トビウオ	トビウオ	和	雑談	35-39歳	女性	東京都	東京都
		15010	IC02_理奈子	の	ノ	の		助詞-格助詞			ノ	ノ	和	雑談	35-39歳	女性	東京都	東京都
		15020	IC02_理奈子	出汁	ダシ	出し		名詞-普通名詞-一般			ダシ	ダシ	和	雑談	35-39歳	女性	東京都	東京都
		30	IC02_理奈子	と	ト	と		助詞-格助詞			ト	ト	和	雑談	35-39歳	女性	東京都	東京都
		15040	IC02_理奈子	か	カ	か		助詞-副助詞			カ	カ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26350	15050	IC02_理奈子	売っ	ウル	売る		動詞-一般	五段-ラ行	連用形-促音便	ウッ	ウッ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26370	15060	IC02_理奈子	て	テル	てる		助動詞	下一段-タ行	連用形-一般	テ	テ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26380	15070	IC02_理奈子	た	タ	た		助動詞	助動詞-タ	終止形-一般	タ	タ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26390	15080	IC02_理奈子	けど	ケレド	けれど		助詞-接続助詞			ケド	ケド	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26410	15090	IC02_理奈子	ね	ネ	ね		助詞-終助詞			ネ	ネ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26420	15100	IC04_ばあば	トビウオ	トビウオ	飛び魚		名詞-普通名詞-一般			トビウオ	トビウオ	和	雑談	65-69歳	女性	奈良県	東京都
T016_002	26460	15110	IC04_ばあば	の	ノ	の		助詞-格助詞			ノ	ノ	和	雑談	65-69歳	女性	奈良県	東京都
T016_002	26470	15120	IC04_ばあば	出汁	ダシ	出し		名詞-普通名詞-一般			ダシ	ダシ	和	雑談	65-69歳	女性	奈良県	東京都
T016_002	26490	15130	IC04_ばあば	おいしい	オイシイ	美味しい		形容詞-一般	形容詞	終止形-一般	オイシー	オイシー	和	雑談	65-69歳	女性	奈良県	東京都
T016_002	26530	15140	IC04_ばあば	ね	ネ	ね		助詞-終助詞			ネ	ネ	和	雑談	65-69歳	女性	奈良県	東京都
T016_002	26540	15150	IC02_理奈子	で	デ	で		接続詞			デ	デ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26550	15160	IC02_理奈子	ね	ネ	ね		助詞-終助詞			ネ	ネ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26560	15170	IC02_理奈子	すごい	スゴイ	凄い		形容詞-一般	形容詞	連体形-一般	スゴイ	スッゴイ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26590	15180	IC02_理奈子	高かっ	タカイ	高い		形容詞-一般	形容詞	連用形-促音便	タカカッ	タカカッ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26620	15190	IC02_理奈子	た	タ	た		助動詞	助動詞-タ	連体形-一般	タ	タ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26630	15200	IC02_理奈子	の	ノ	の		助詞-終助詞			ノ	ノ	和	雑談	35-39歳	女性	東京都	東京都
T016_002	26640	15210	IC04_ばあば	高い	タカイ	高い		形容詞-一般	形容詞	終止形-一般	タカイ	タカイ	和	雑談	65-69歳	女性	奈良県	東京都
T016_002	26660	15220	IC04_ばあば	ね	ネ	ね		助詞-終助詞			ネ	ネ	和	雑談	65-69歳	女性	奈良県	東京都
T016_002	26670	15230	IC03_謙一	ついで	ツイデ	序で		名詞-普通名詞-一般			ツイデ	ツイデ	和	雑談	10-14歳	男性	東京都	東京都
T016_002	26700	15240	IC03_謙一	に	ニ	に		助詞-格助詞			ニ	ニ	和	雑談	10-14歳	男性	東京都	東京都
T016_002	26710	15250	IC03_謙一	俺っち	オレッチ	俺っち		代名詞			オレッチ	オレッチ	和	雑談	10-14歳	男性	東京都	東京都
T016_002	26740	15260	IC03_謙一	も	モ	も		助詞-係助詞			モ	モ	和	雑談	10-14歳	男性	東京都	東京都
T016_002	26750	15270	IC03_謙一	お	オ	御		接頭辞			オ	オ	和	雑談	10-14歳	男性	東京都	東京都
T016_002	26760	15280	IC03_謙一	願い	ネガウ	願う		動詞-非自立可能	五段-ワア行	連用形-一般	ネガイ	ネガイ	和	雑談	10-14歳	男性	東京都	東京都
T016_002	26780	15290	IC03_謙一	し	スル	為る		動詞-非自立可能	サ行変格	連用形-一般	シ	シ	和	雑談	10-14歳	男性	東京都	東京都
T016_002	26790	15300	IC03_謙一	やす	ヤンス	やんす		助動詞	助動詞-ンス	終止形-一般	ヤス	ヤース	和	雑談	10-14歳	男性	東京都	東京都

オンライン検索システムの限界

- 文脈を把握しづらい（特に会話）
- 音声の部分的視聴しかできない
- 形態論情報とメタ情報しか表示できない

公開方法2：生データの公開（NUCC・CWPC以外）

- 転記テキストを利用して文脈を追うことができる
- 音声・映像（CEJC,CEJC-CHILDのみ）を利用可能

  ※学習者コーパスは無償だがMP4，それ以外は有償だがWAV
- 形態論情報以外のアノテーションを利用可能

  ※学習者コーパス（無償）はMP4，それ以外（有償）はWAV
- 複数のアノテーションを統合したRDB等を提供

  ※CSJ, CEJC のみ

2つのコーパスの紹介

- 日本語話し言葉コーパス（CSJ）
- 日本語日常会話コーパス（CEJC）

日本語話し言葉コーパス
Corpus of Spontaneous Japanese, CSJ

『日本語話し言葉コーパス』
Corpus of Spontaneous Japanese, CSJ

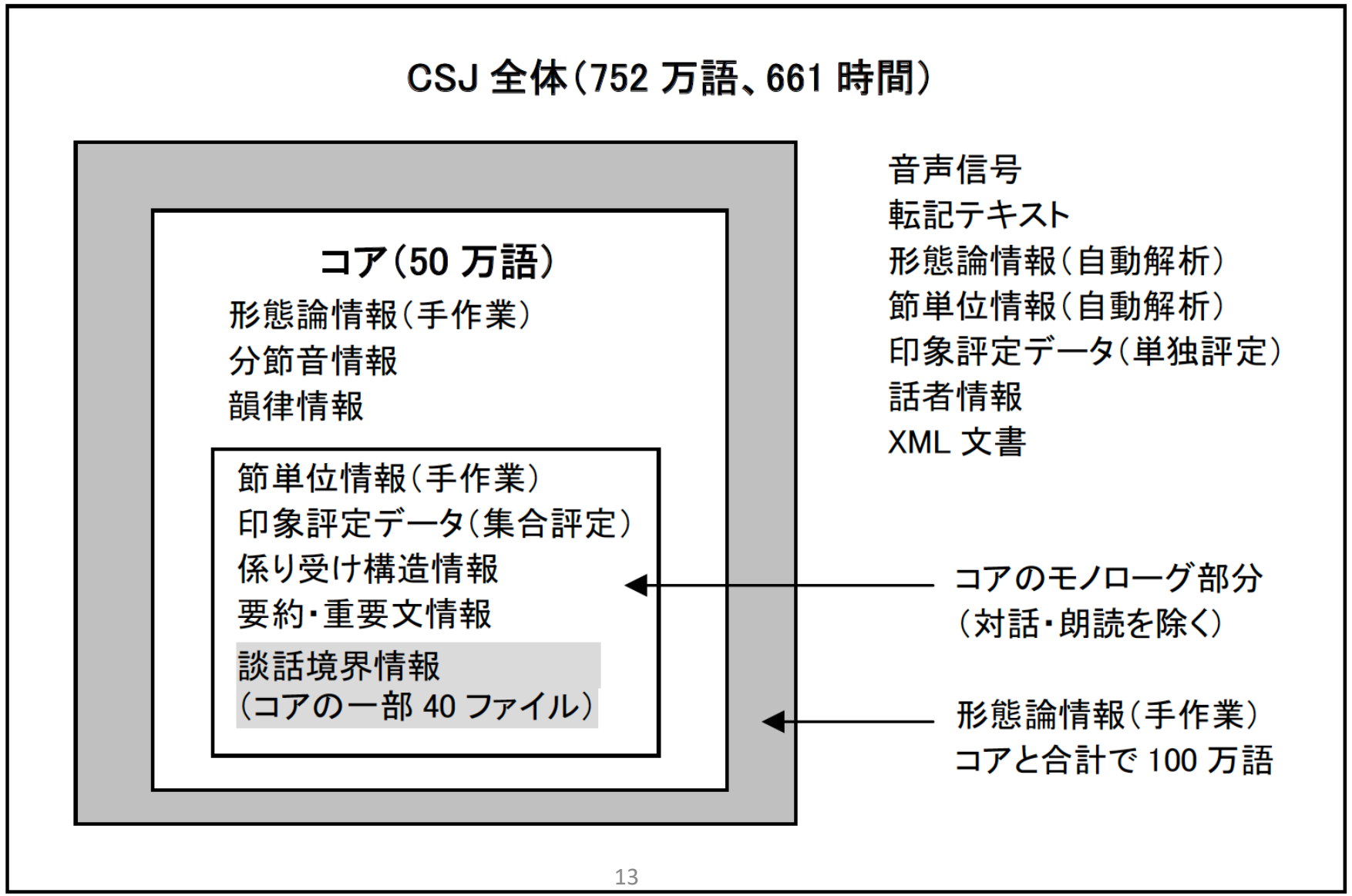
- プロジェクト：文科省科学技術振興調整費開放的融合研究制度研究課題
  「話し言葉の言語的・パラ言語的構造の解析に基づく『話し言葉工学』の構築」

- 構築期間：1999～2003年度
- 対　象：自発音声（独話・会話）・再朗読・朗読
- 規　模：660時間・750万語
- 開発目的：
  - ➢自発音声の音声認識用言語モデル・音響モデルの学習データ
  - ➢自発音声の言語学的研究のリソース

音声のタイプ

収録期間：1999～2003

ＣＳＪの設計


CSJ 全体（752 万語、661 時間）
コア（50 万語）
形態論情報（手作業）
分節音情報
韻律情報
節単位情報（手作業）
印象評定データ（集合評定）
係り受け構造情報
要約・重要文情報
談話境界情報
（コアの一部 40 ファイル）
音声信号
転記テキスト
形態論情報（自動解析）
節単位情報（自動解析）
印象評定データ（単独評定）
話者情報
XML 文書
コアのモノローグ部分
（対話・朗読を除く）
形態論情報（手作業）
コアと合計で 100 万語

音声の文字化

00518.251 00520.292 L: 五月からの三か月って(F えー)
00520.801 00522.865 L: 御存じかどうか分からないです分かりませんが
00523.067 00523.663 L: 税金
00524.138 00525.306 L: のね額を決める
00525.887 00527.615 L: (F えー)標(笑 準報酬月額)って
00528.199 00529.256 L: いうのをね決める
00529.505 00529.915 L: <笑>
00530.107 00531.096 L: 対象になるんですよ
00532.095 00532.457 L: <笑>
00532.489 00533.235 L: つまりですね<H>
00533.825 00534.757 L: (W モク;僕)らその月
00535.346 00536.181 L: 幾らぐらいかな
00536.460 00537.434 L: 四十万とか
00537.967 00538.499 L: (F ん)(F まー)
00538.502 00539.723 L: <笑>
00539.963 00542.662 L: そのその手取りは(F えー)三十なっても
00543.031 00545.518 L: 貰えないよとかって(笑 その工場実習で)
00545.842 00548.039 L: その後現場の人に言われたぐらいの額をですね
00548.639 00553.683 L: 貰った貰ってもう大喜びしてたんですけど次の年の税金が非常に(笑 高くて)

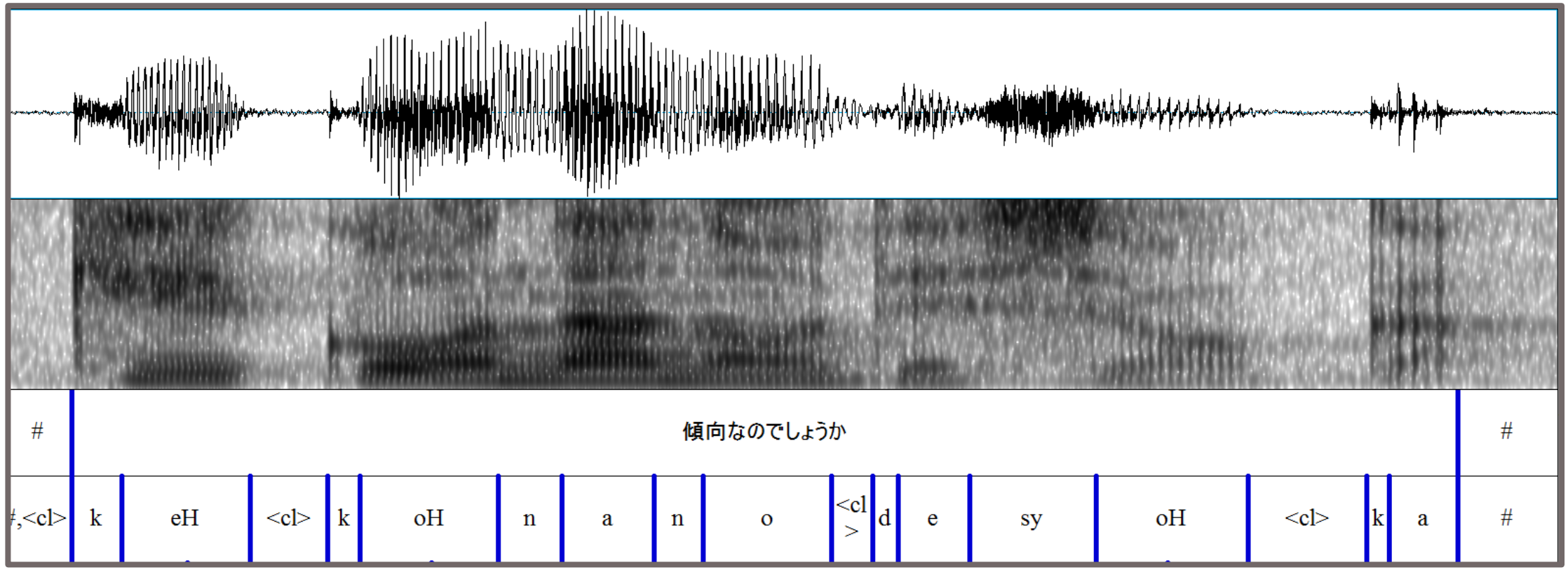
形態論情報

- 提供する形態論情報の種類：短単位情報・長単位情報
- 注意事項
  - 中納言版は現時点では短単位のみ検索可能
  - USB版と中納言版では単位の粒度・品詞体系が異なる

例)	USB版	中納言版（UniDic体系）
短単位の粒度	|オレンジケーキ| |オレンジ色|	|オレンジ|ケーキ| |オレンジ|色|
短単位の品詞	名詞 動詞	名詞-普通名詞-サ変可能 動詞-非自立可能

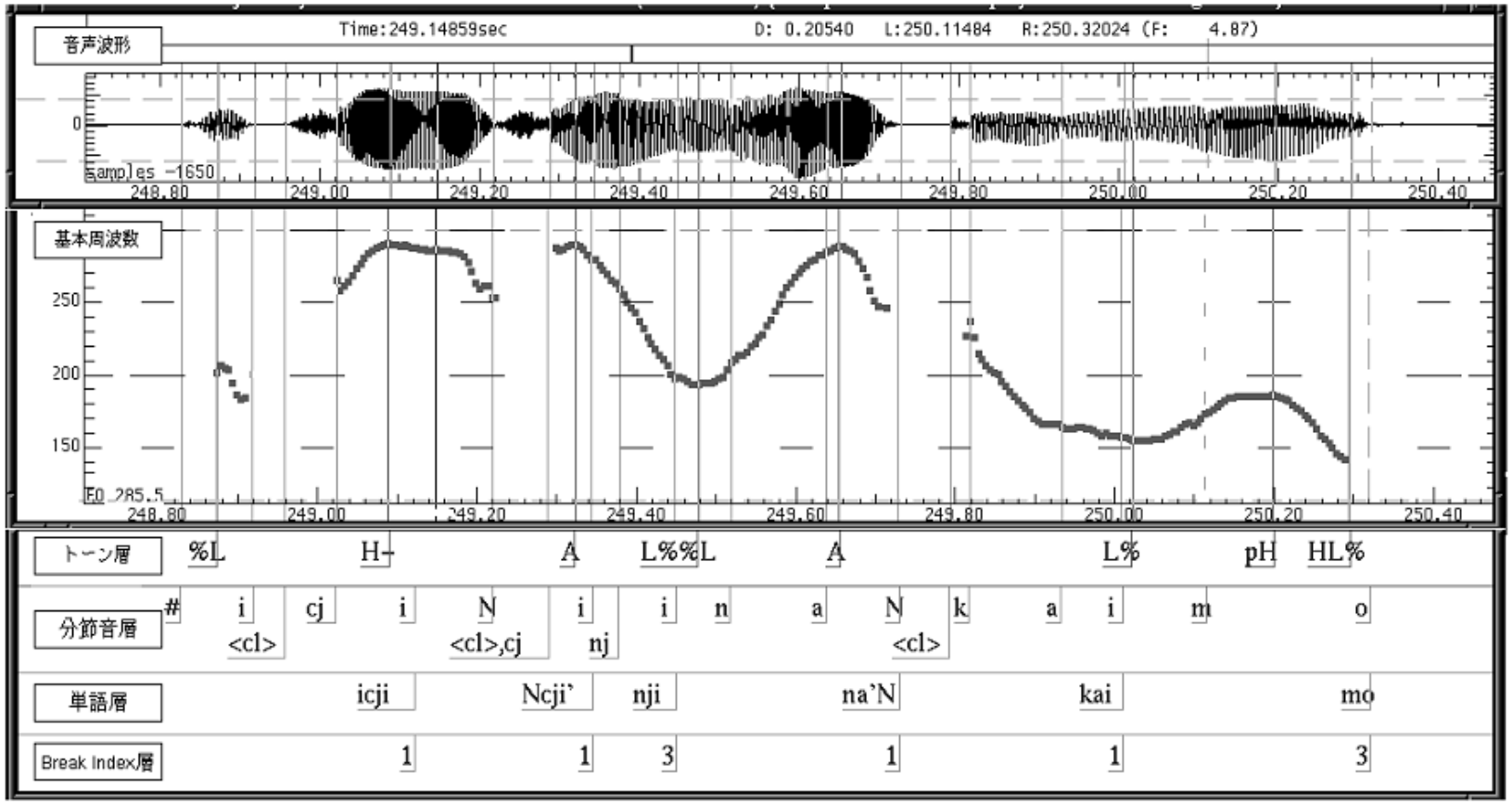
分節音情報

- 母音・子音等の種類とその時刻に関する情報



韻律情報

- J-ToBI（Venditti)を自発音声用に拡張したX-JToBIに準拠しラベリング
- 基本周波数曲線を音韻的なイベントとして記号化



NINJAL
National Institute for Japanese Language and Linguistics

節単位情報

- 節境界の位置と種類（例：ケレドモ、ノデ）を特定
- 節境界を、境界直後の構造的な切れ目の大きさ（主節に対する従属の程度）の観点から、以下の3種類に分類

【絶対境界】いわゆる文末に相当する境界
【強境界】　切れ目の度合が強い（主節に対し従属度が弱い）境界
【弱境界】　切れ目の度合が弱い（主節に対し従属度が強い）境界

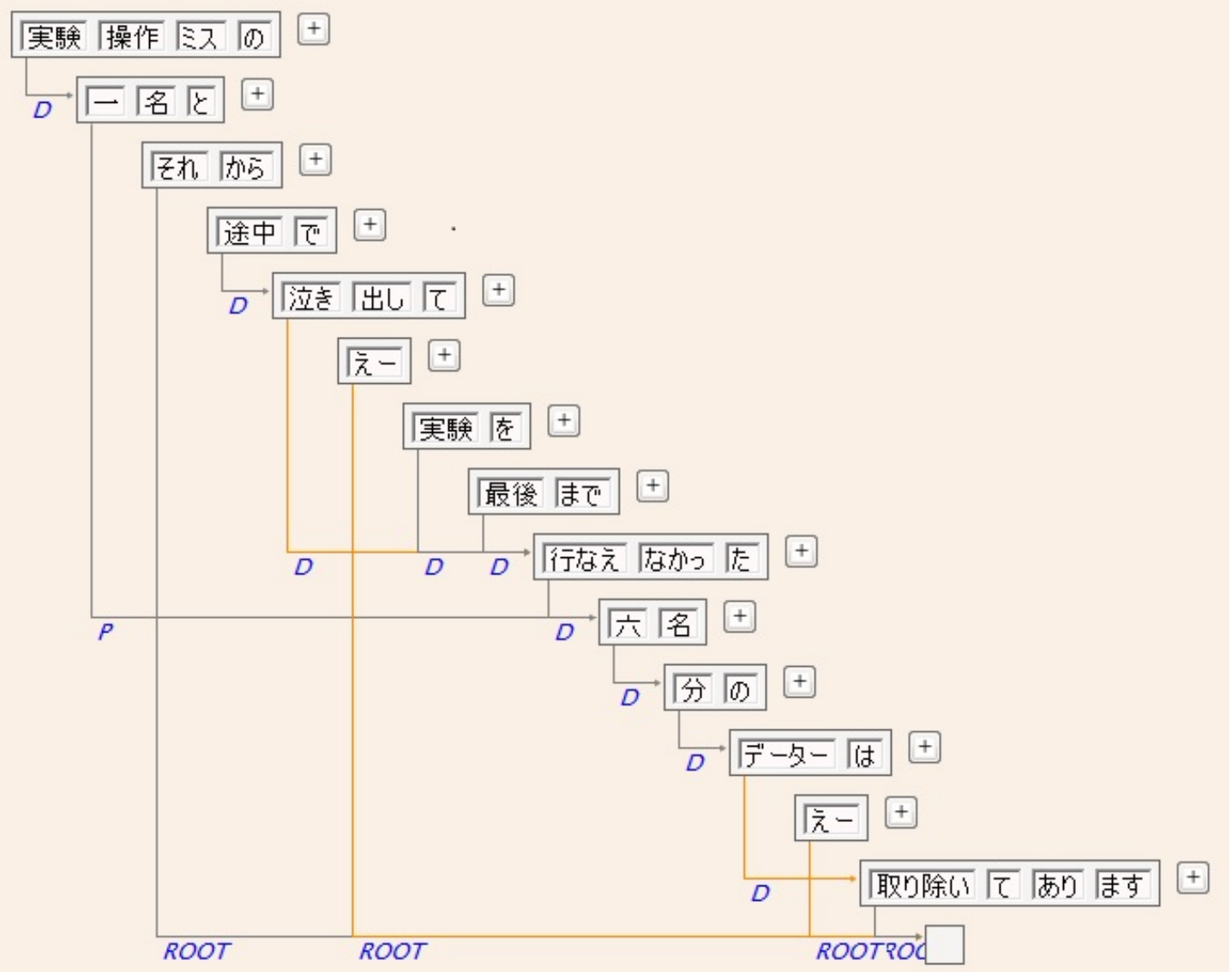
➢絶対境界か強境界で分割される単位を「節単位」と認定
➢体言止めや非流暢現象などは人手で境界を修正

※検索システム中納言でも「文」に代わる単位として利用

節単位情報

発話例	節境界ラベル	人手修正
やっぱり(F えー)高校時代とかもうどうしてもこう 生きるって何だろうとか人間て何だろうみたいなの(D2 を) を(F えー)悩むと思うんですが	/並列節ガ/	
それの何か答えを見つけたような(0.129)気がしました人間失格を読んで	<テ節>	倒置
(F ま)暗い話ですけれども	/並列節ケレドモ/	
(F え)そこが非常に心に残っております	[文末]	
それと(F えー)もう一人 村上龍		体言止

係り受け情報



文節間の係り受け

単独印象評定情報　（判定者１名・全体対象）

- 段階評定(5段階で評定)

講演の自発性、発話スピード、発音の明瞭さ、方言の多少、
発話スタイル

- 評定語選択式(該当するものにチェック)

たどたどしい、流暢な、単調な、表情ゆたかな、自信のある、
自信のない、優しい、落ち付いた、落ち付きのない、
いらいらした、緊張した、リラックスした、重厚な、軽薄な、
若々しい、年寄じみた、元気のある、元気のない、
聞き取りやすい、聞き取りにくい、生意気な、尊大な、等

集合印象評定情報　（判定者１０名・コアのみ対象）

◆ 講演音声評定尺度

好悪	A01	好きな—嫌いな
	A02	心地よい—不快な
	A03	感じの良い—感じの悪い
	A04	親しみやすい—親しみにくい
上手さ	A05	流暢な—たどたどしい
	A06	話し慣れた—話し慣れていない
	A07	なめらかな—しどろもどろな
	A08	上手い—下手な
速さ感	A09	速い—遅い
	A10	スピード感のある—ゆったりした
	A11	せわしげな—のんきな
	A12	落ち着きのない—落ち着きのある
活動性	A13	声の大きい—声の小さい
	A14	力強い—弱々しい
	A15	元気のある—元気のない
	A16	積極的な—消極的な
スタイル	A17	礼儀正しい—無礼な
	A18	まじめな—ふまじめな
	A19	丁寧な—ぞんざいな
	A20	上品な—下品な

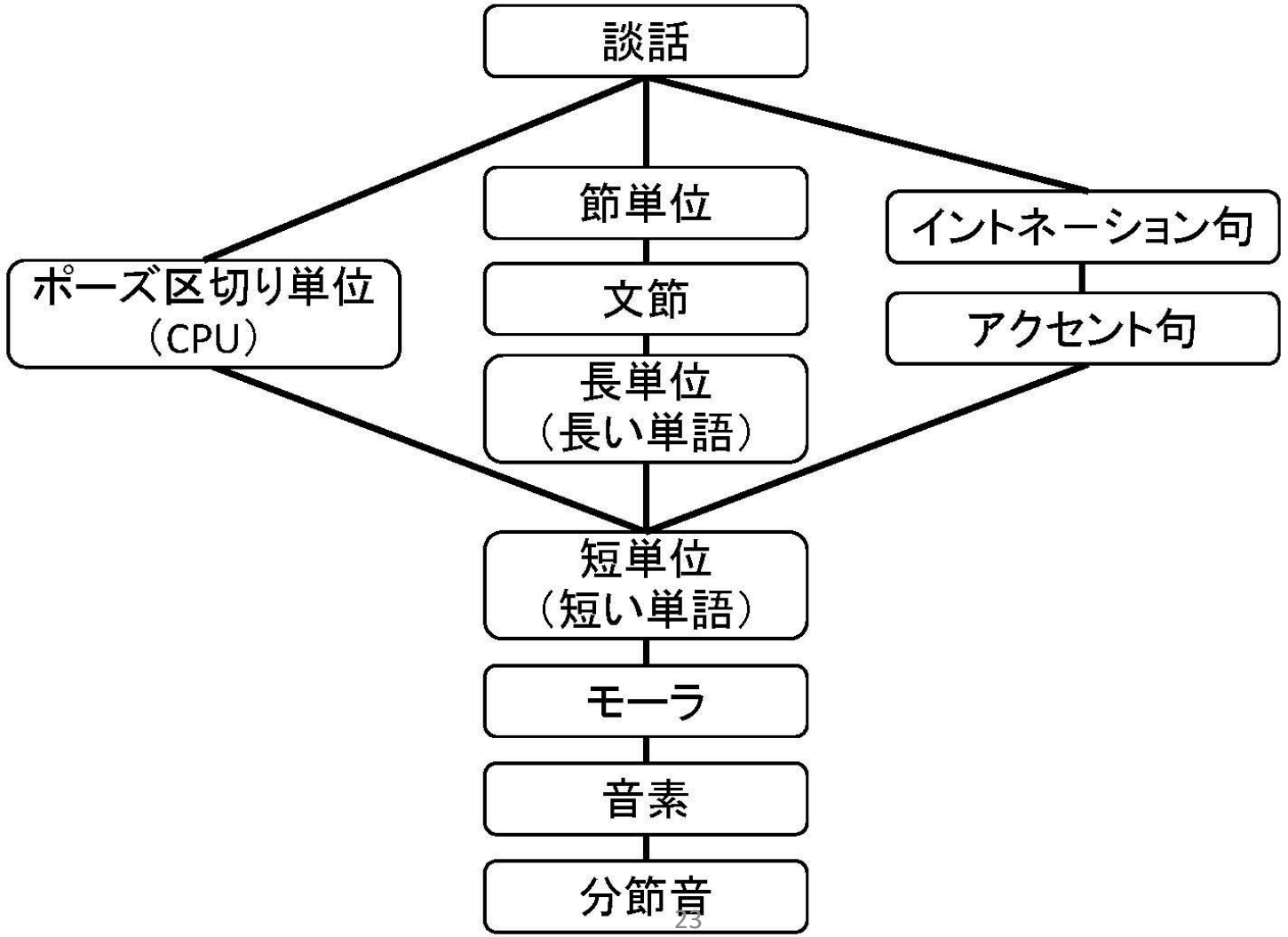
◆ 日本語big five 尺度短縮版

外向性	B01	話し好き
	B02	無口な [*]
	B03	陽気な
	B04	外向的な
情緒不安定性	B05	悩みがち
	B06	不安になりやすい
	B07	心配症
	B08	気苦労の多い
経験への開放性	B09	独創的な
	B10	進歩的
	B11	洞察力のある
	B12	想像力に富んだ
誠実性 (勤勉性)	B13	いい加減な [*]
	B14	ルーズな [*]
	B15	怠惰な [*]
	B16	計画性のある
調和性 (協調性)	B17	温和な
	B18	寛大な
	B19	親切な
	B20	協力的な

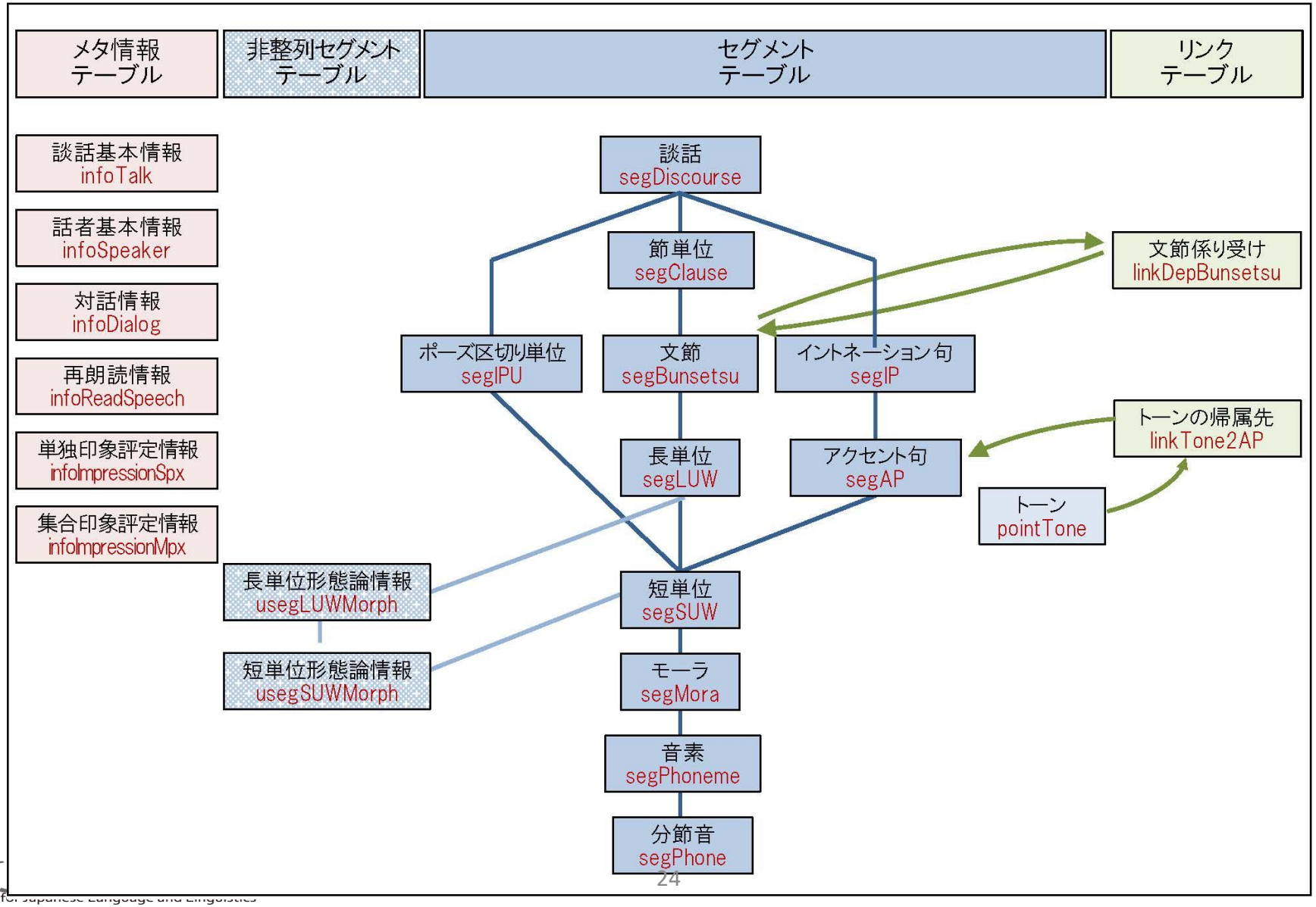
◆ 話し方の印象に関する単項項目

A21	あらたまった—くだけた
A22	きまじめな—奔放な
A23	きちんとした—くつろいだ
A24	甘えた—そっけない
A25	その場で考えて話している—原稿を読み上げている
A26	聞き取りやすい—聞き取りにくい

アノテーションに関わる単位間の関係



CSJ-RDB


メタ情報
テーブル
非整列セグメント
テーブル
セグメント
テーブル
リンク
テーブル
談話基本情報
infoTalk
話者基本情報
infoSpeaker
対話情報
infoDialog
再朗読情報
infoReadSpeech
単独印象評定情報
infoImpressionSpx
集合印象評定情報
infoImpressionMpx
談話
segDiscourse
節単位
segClause
文節係り受け
linkDepBunsetsu
ポーズ区切り単位
segIPU
文節
segBunsetsu
イントネーション句
segIP
トーンの帰属先
linkTone2AP
長単位
segLUW
アクセント句
segAP
トーン
pointTone
長単位形態論情報
usegLUWMorph
短単位
segSUW
短単位形態論情報
usegSUWMorph
モーラ
segMora
音素
segPhoneme
分節音
segPhone

CSJの利用実績


CSJ活用論文数
140
120
100
80
60
40
20
0
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025

CSJの利用実績


商業利用
25
20
15
10
5
0
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025

CSJの利用実績

最近の工学系の利用動向

- 音声認識のベンチマーク（評価セット）
- ファインチューニング
- ドメイン適応
- 誤り解析

など

日本語日常会話コーパス
Corpus of Everyday Japanese Conversation, CEJC

『日本語日常会話コーパス』
Corpus of Everyday Japanese Conversation, CEJC

- プロジェクト：大規模日常会話コーパスに基づく話し言葉の多角的研究
- 構築期間：2016～2021年度
- 対　象：日常生活の中で生じるリアルな会話
- 規　模：200時間

データ規模

時間	200時間
会話数	577会話
話者数（延べ）	1675名
話者数（異なり）	862名
語数	240万語

構築の背景

主要な日本語の会話コーパス

- 話題・収録の状況：
  - ✓ 実験環境，集まってもらった状況での雑談
  - ✓ 日常生活で実際に行われた会話（少）
- 会話の種類：
  - ✓ 偏った種類の会話・話者（例：大学生の雑談、電話会話など）
  - ✓ 多様な種類の会話（少）
- 公開データ：
  - ✓ 転記テキストだけ
  - ✓ 転記テキスト ＋　音声データ
  - ✓ 転記テキスト ＋　音声データ　＋　映像データ （少）
- アノテーション：
  - ✓豊富なアノテーションの付いた会話コーパス （少）

日常会話コーパスの収録法

■ 個人密着法　　185時間

- ✓ 性別・年齢の観点からバランスを考慮して選別された協力者に収録依頼
- ✓ （男女×年齢5世代×各4人=40人、職業偏らないよう配慮）
- ✓ 機材機器等を3か月ほど貸し出し，協力者の日常生活で自発的に生じるリアルな会話を記録（1協力者あたり平均約15時間収録）
- ✓ コーパス構成比や倫理的問題等を考慮してコーパスに含める会話を選別
  - 1協力者あたり約4-5時間を選別，計160-200時間（目安）

■ 特定場面法　　15時間

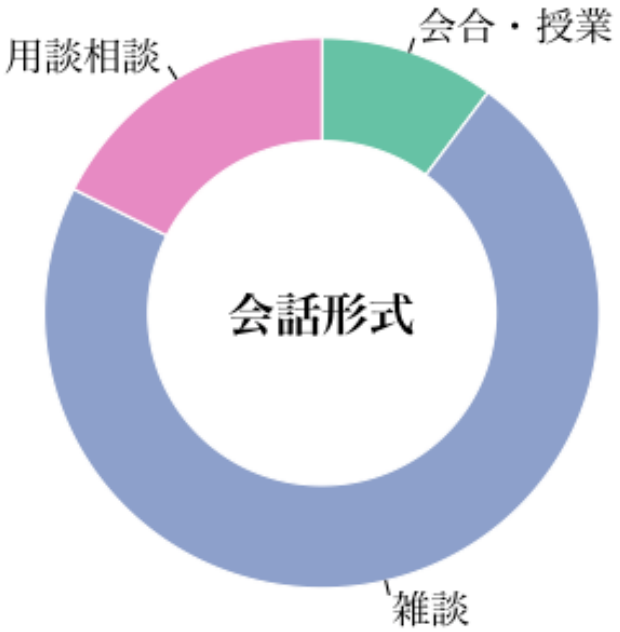
個人密着法では収録の難しい場面

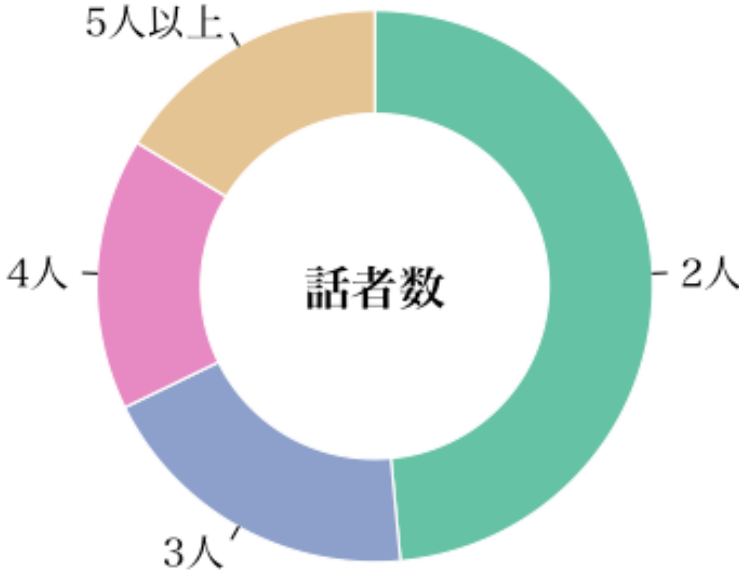
- ✓ 仕事場面での会議・会合
- ✓ 未成年者（中高生）中心の会話

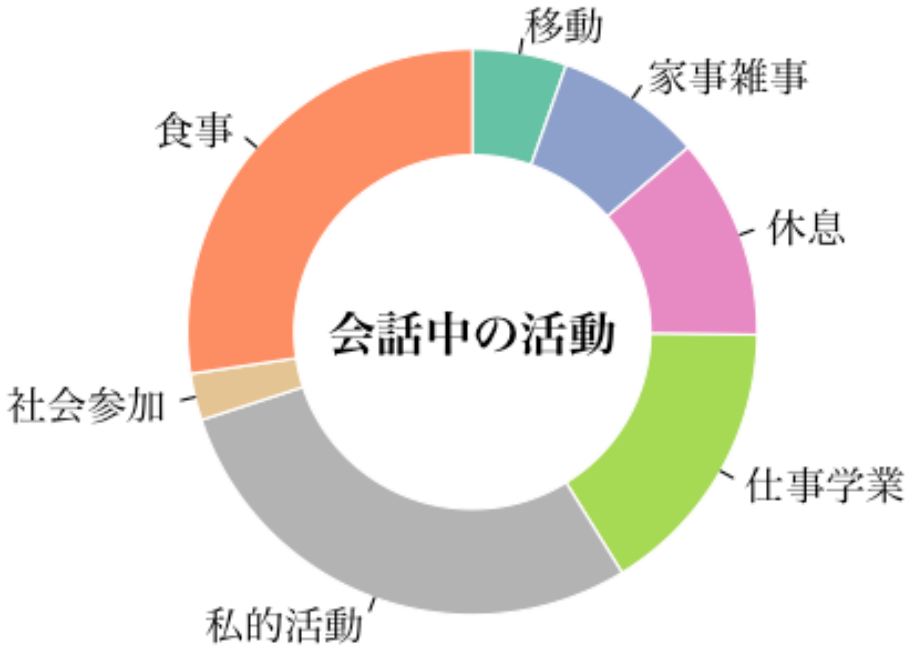
既存のコーパスの利用可能性の向上①
話者間の関係性に関する情報

年代	男性		女性	
	職業・職種	時間	職業・職種	時間
20代	学生	4.2h	学生	6.0h
	学生	4.3h	大学生	4.4h
	先生	3.7h	会社員・公務員等	4.2h
	先生	5.5h	会社員・公務員等	4.0h
30代	自営業・自由業	5.6h	会社員・公務員等	5.0h
	会社員・公務員等	4.6h	専業主婦	5.6h
	会社員・公務員等	4.7h	自営業・自由業	5.4h
	会社員・公務員等	3.1h	自営業・自由業	4.8h
40代	会社員・公務員等	3.6h	会社員・公務員等	4.5h
	自営業・自由業	4.8h	パート・アルバイト	4.8h
	先生	3.9h	パート・アルバイト	5.0h
	会社員・公務員等	5.0h	自営業・自由業	4.4h
50代	会社員・公務員等	6.0h	会社員・公務員等	4.2h
	会社員・公務員等	4.6h	自営業・自由業	4.6h
	先生	4.2h	自営業・自由業	4.6h
	先生	4.2h	会社員・公務員等	4.5h
60歳～	先生	4.3h	専業主婦	5.1h
	定年退職	5.8h	自営業・自由業	4.4h
	会社員・公務員等	4.8h	会社員・公務員等	4.2h
	定年退職	4.6h	自営業・自由業	4.3h

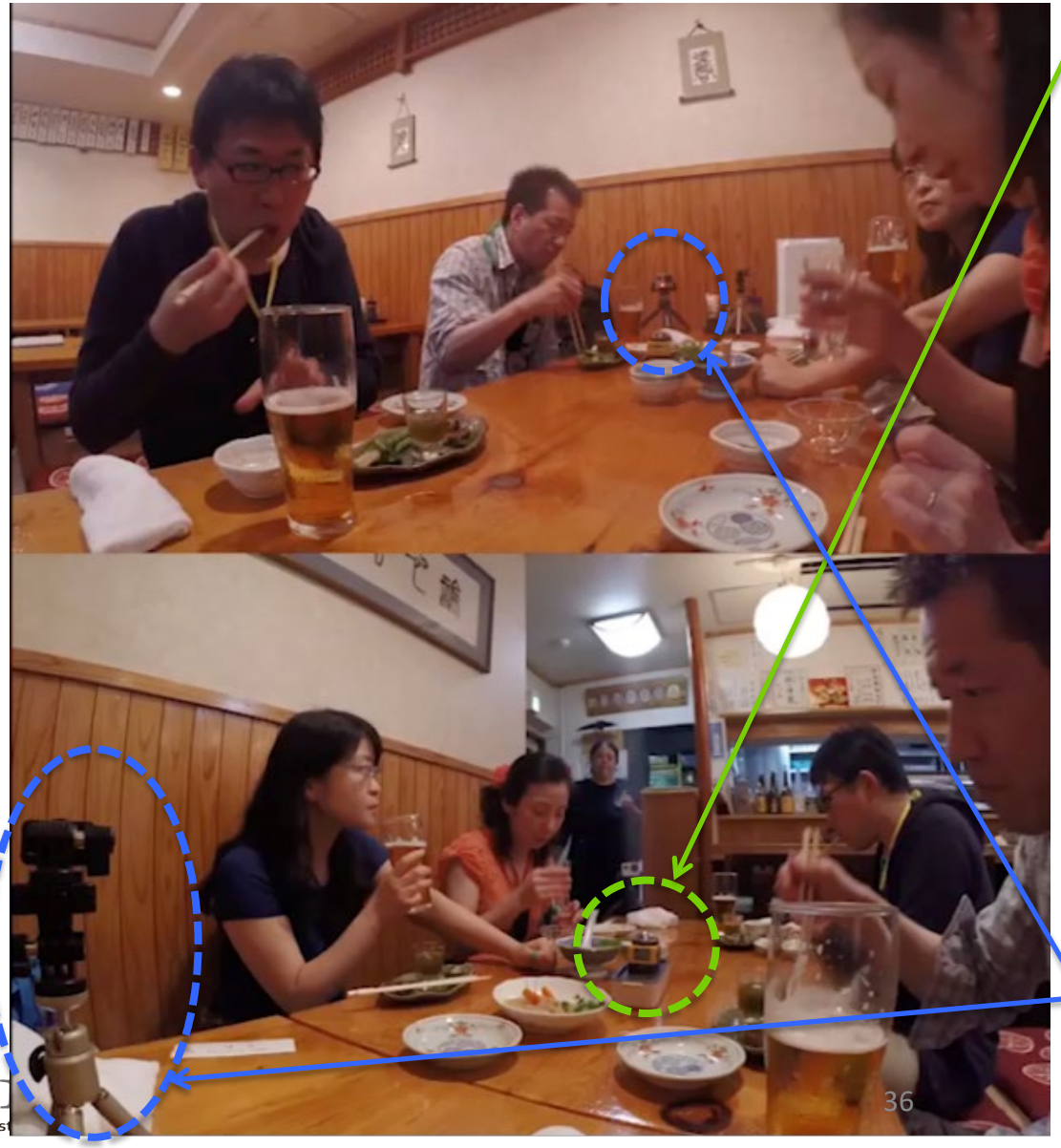
既存のコーパスの利用可能性の向上①
話者間の関係性に関する情報







映像の収録（基本）



Kodak PIXPRO SP360 4K

会話者の中心に360度撮影可能なカメラを配置

GoPro Hero3+

会話を俯瞰的に記録するカメラを1～2台配置

音声の収録

話者ごとにICレコーダーを首から下げた
フォルダーに入れて装着し、
当該話者の音声を中心に収録

個人ICレコーダー

Sony ICD-SX734

中央ICレコーダー

Sony ICD-SX1000

会話の場の中央に置いて
会話全体を収録

提供データ

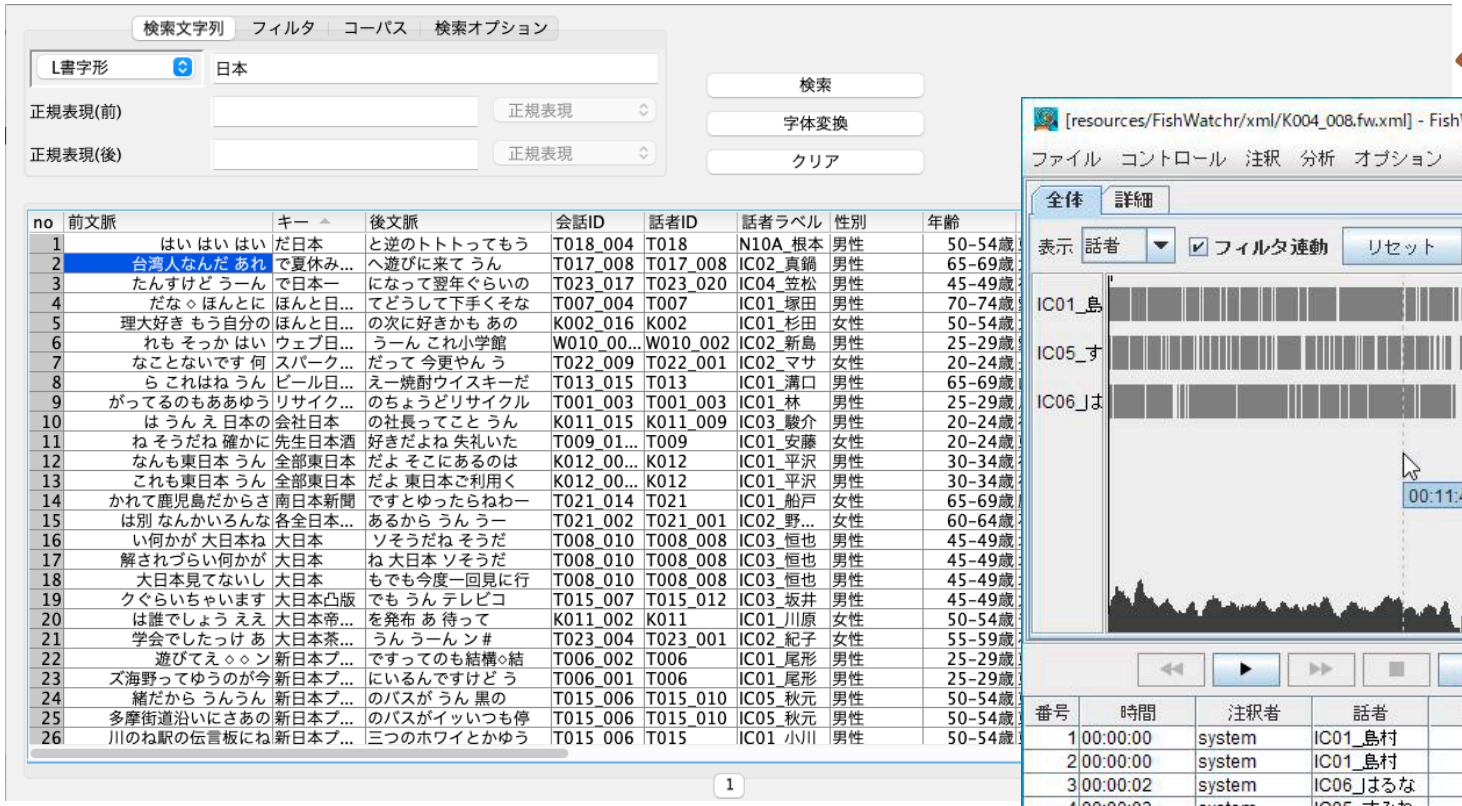
CEJC全体　200時間

映像・音声データ
転記テキスト
[人手修正]　形態論情報（短単位情報）
[自動付与]　形態論情報（長単位情報）

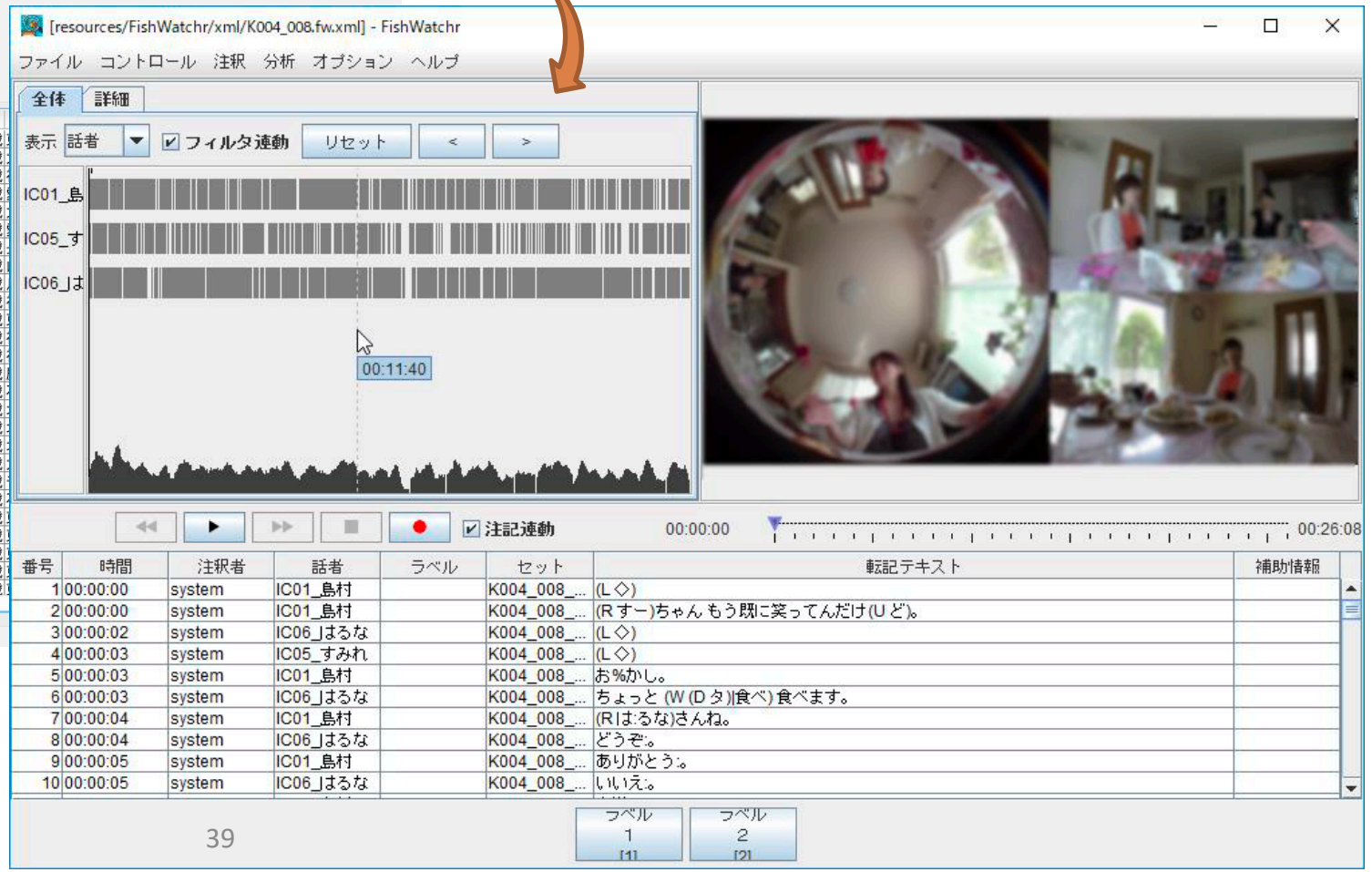
コア　20時間

[人手修正]　形態論情報（長単位情報）
係り受け情報
[人手付与]　談話行為情報・韻律情報

検索システム「ひまわり」+動画再生



no	前文脈	キー	後文脈	会話ID	話者ID	話者ラベル	性別	年齢
1	はい はい はい	だ日本	と逆のトトトってもう	T018_004	T018	N10A_根本	男性	50-54歳
2	台湾人なんだ あれ	で夏休み...	へ遊びに来て うん	T017_008	T017_008	IC02_真鍋	男性	65-69歳
3	たんすけど うーん	で日本ー	になって翌年ぐらいの	T023_017	T023_020	IC04_笠松	男性	45-49歳
4	だな ◇ ほんとに	ほんと日...	てどうして下手くそな	T007_004	T007	IC01_塚田	男性	70-74歳
5	理大好き もう自分の	ほんと日...	の次に好きかも あの	K002_016	K002	IC01_杉田	女性	50-54歳
6	れも そっか はい	ウェブ日...	うーん これ小学館	W010_00...	W010_002	IC02_新島	男性	25-29歳
7	なことないです 何	スパーク...	だって 今更やん う	T022_009	T022_001	IC02_マサ	女性	20-24歳
8	ら これはね うん	ビール日...	えー焼酎ウイスキーだ	T013_015	T013	IC01_溝口	男性	65-69歳
9	がってるのもああゆう	リサイク...	のちょうどリサイクル	T001_003	T001_003	IC01_林	男性	25-29歳
10	は うん え 日本の	会社日本	の社長ってこと うん	K011_015	K011_009	IC03_駿介	男性	20-24歳
11	ね そうだね 確かに	先生日本酒	好きだよね 失礼いた	T009_01...	T009	IC01_安藤	女性	20-24歳
12	なんも東日本 うん	全部東日本	だよ そこにあるのは	K012_00...	K012	IC01_平沢	男性	30-34歳
13	これも東日本 うん	全部東日本	だよ 東日本ご利用く	K012_00...	K012	IC01_平沢	男性	30-34歳
14	かれて鹿児島だからさ	南日本新聞	ですとゆったらねー	T021_014	T021	IC01_船戸	女性	65-69歳
15	は別 なんかいろんな	各全日本...	あるから うん うー	T021_002	T021_001	IC02_野...	女性	60-64歳
16	い何かが 大日本ね	大日本	ソそうだね そうだ	T008_010	T008_008	IC03_恒也	男性	45-49歳
17	解されづらい何かが	大日本	ね 大日本 ソそうだ	T008_010	T008_008	IC03_恒也	男性	45-49歳
18	大日本見てないし	大日本	もでも今度一回見に行	T008_010	T008_008	IC03_恒也	男性	45-49歳
19	クぐらいちゃいます	大日本凸版	でも うん テレビコ	T015_007	T015_012	IC03_坂井	男性	45-49歳
20	は誰でしょう ええ	大日本帝...	を発布 あ 待って	K011_002	K011	IC01_川原	女性	50-54歳
21	学会でしたっけ あ	大日本茶...	うん うーん ン#	T023_004	T023_001	IC02_紀子	女性	55-59歳
22	遊びてえ ◇ ◇ ン	新日本プ...	ですってのも結構◇結	T006_002	T006	IC01_尾形	男性	25-29歳
23	ズ海野ってゆうのが今	新日本プ...	にいるんですけど う	T006_001	T006	IC01_尾形	男性	25-29歳
24	緒だから うんうん	新日本プ...	のバスが うん 黒の	T015_006	T015_010	IC05_秋元	男性	50-54歳
25	多摩街道沿いにさあの	新日本プ...	のバスがイッいつも停	T015_006	T015_010	IC05_秋元	男性	50-54歳
26	川のね駅の伝言板にね	新日本プ...	三つのホワイとかゆう	T015_006	T015	IC01_小川	男性	50-54歳



番号	時間	注釈者	話者	ラベル	セット	転記テキスト	補助情報
1	00:00:00	system	IC01_島村		K004_008_...	(L ◇)	
2	00:00:00	system	IC01_島村		K004_008_...	(R すー)ちゃん もう既に笑ってんだけ(U ど)。	
3	00:00:02	system	IC06_はるな		K004_008_...	(L ◇)	
4	00:00:03	system	IC05_すみれ		K004_008_...	(L ◇)	
5	00:00:03	system	IC01_島村		K004_008_...	お%かし。	
6	00:00:03	system	IC06_はるな		K004_008_...	ちょっと (W (D タ)|食べ) 食べます。	
7	00:00:04	system	IC01_島村		K004_008_...	(R は:るな)さんね。	
8	00:00:04	system	IC06_はるな		K004_008_...	どうぞ。	
9	00:00:05	system	IC01_島村		K004_008_...	ありがとう:。	
10	00:00:05	system	IC06_はるな		K004_008_...	いいえ:。	

談話行為情報

ISO 24617-2 をベースに日常会話用に整備した基準に基づき、
発話単位ごとに人手で付与

レベル1タグ：基本的な談話機能

タスク	情報提供・情報要求・ 依頼系・申し出・ 注意獲得・独り言 …
社会的付き合い管理	挨拶・謝罪・感謝・ 謝罪への対処 …
フィードバック FB	FB肯定（あいづち） FB了承 FB反復 FB語彙的反応 …

レベル2タグ：該当する場合

修復	修復開始・修復操作 …
談話構造化	談話開始・談話終了 …
順番管理	順番取得・順番維持 …

依存関係

予測的	（いわゆる隣接ペア）
遡及的	（フィードバックなど）
外部予測的	（第1ペアのない隣接ペア）

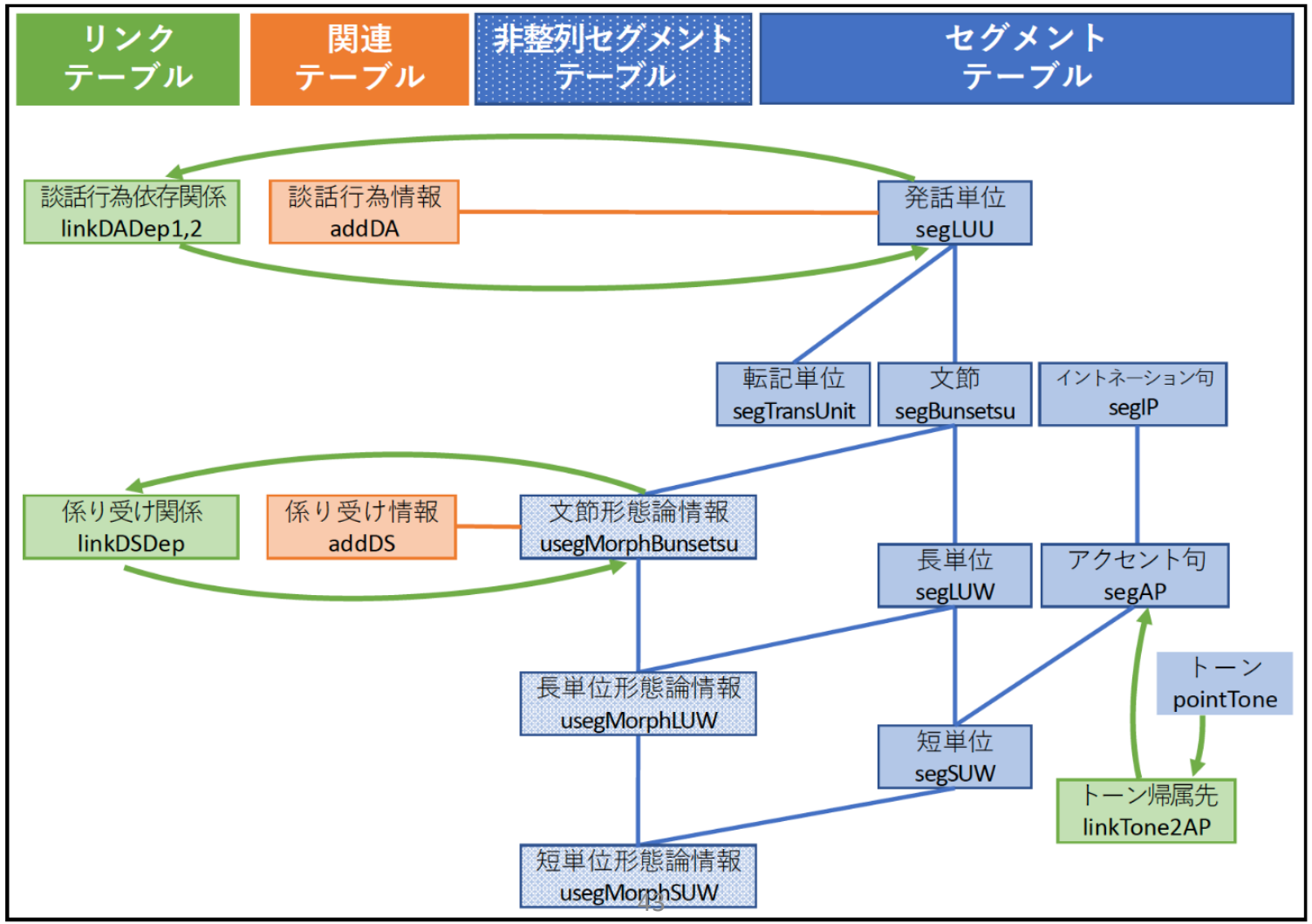
会話に関するメタ情報

話者数	主たる話者の数　（一時的に話に加わった話者などは除く）
形式	主たる会話の形式　（会合中に雑談が生じるなど複数の形式が関わりうるが主たる形式を１つ認定）
場所	会話が行われた場所
活動	会話中の活動　（何をしながら会話をしていたか、最大２つの活動を認定）
話者間の関係性	話者間の関係　（組合せで複数の関係性を認定することあり）
収録年	会話を収録した年
収録法	収録法の別
備考	協力者へのヒアリングなどを通して得られた補足情報

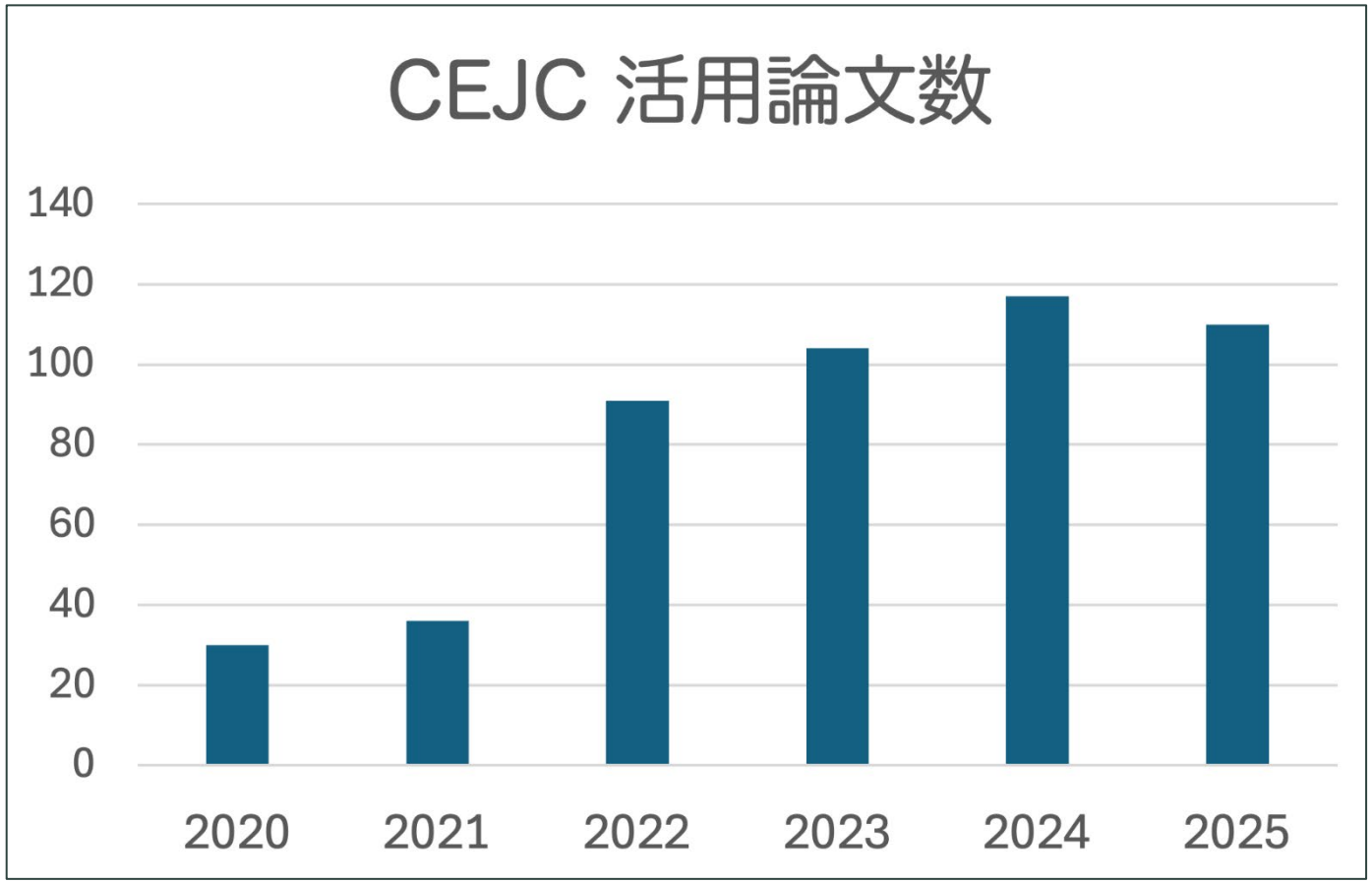
話者に関するメタ情報

話者ID	話者を一意に同定する固有のID
話者ラベル	話者に与えられた仮名のラベル
年齢	収録当時の年齢（5歳刻み）
性別	話者の性別
出身地	話者の出身地（都道府県レベル・外国の場合は国レベル）
居住地	話者の出身地（都道府県レベル・外国の場合は国レベル）
職業	話者の職業
協力者からみた関係性	協力者からみた会話相手との関係性（個人密着法による収録の場合のみ）
備考	協力者へのヒアリングなどを通して得られた補足情報

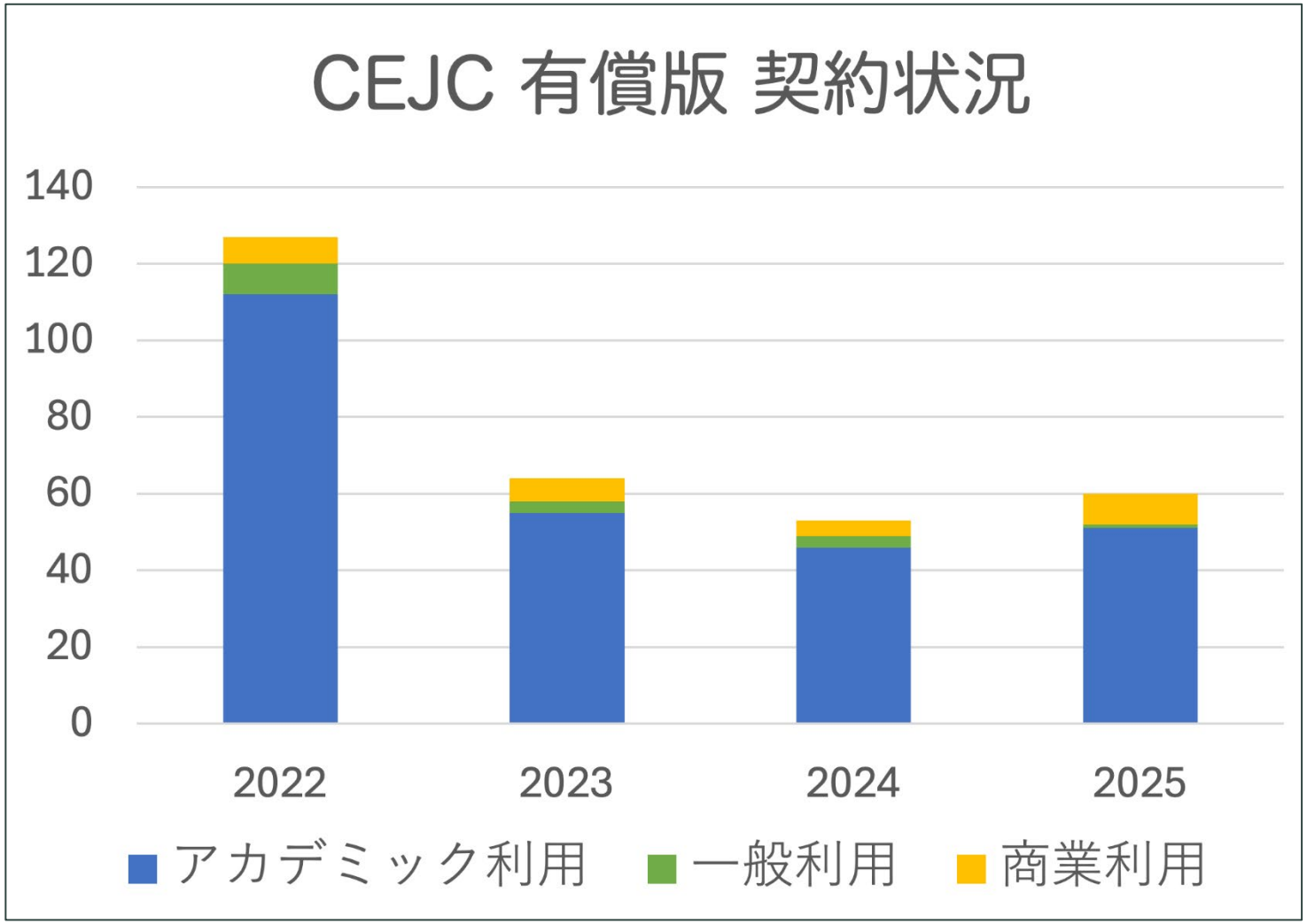
CEJC-RDB


リンク
テーブル
関連
テーブル
非整列セグメント
テーブル
セグメント
テーブル
談話行為依存関係
linkDADep1,2
談話行為情報
addDA
発話単位
segLUU
転記単位
segTransUnit
文節
segBunsetsu
イントネーション句
segIP
係り受け関係
linkDSDep
係り受け情報
addDS
文節形態論情報
usegMorphBunsetsu
長単位
segLUW
アクセント句
segAP
トーン
pointTone
長単位形態論情報
usegMorphLUW
短単位
segSUW
トーン帰属先
linkTone2AP
短単位形態論情報
usegMorphSUW

CEJCの利用実績


CEJC 活用論文数
140
120
100
80
60
40
20
0
2020
2021
2022
2023
2024
2025

CEJCの利用実績


CEJC 有償版 契約状況
140
120
100
80
60
40
20
0
2022
2023
2024
2025
アカデミック利用
一般利用
商業利用

CSJの利用実績

人文系の研究動向

- コミュニケーション研究
- 語用論・談話研究
- 社会言語学など

理工系の研究動向

- 音声認識・音声処理
- 対話処理・対話理解
- マルチモーダル研究

その他の領域の研究動向

- 自閉スペクトラム症者の言語特徴の分析

マルチモーダル研究
動画を対象とした物体検出・追跡

自閉スペクトラム症（ASD）に関する研究

国立障害者リハビリテーションセンター研究所ほかとの共同研究

- ASDは社会的コミュニケーション・社会的相互作用に困難さが見られることが中核症状の１つ

- ASDの幼児は終助詞をほとんど使わない（佐竹・小林 1987など）
- 成人ASD者は終助詞の使用頻度がTD者と異なる（Naoe et al. 2021）

→ １例（綿巻, 1997）～3例（佐竹・小林, 1987）の事例研究
ASD一般の問題とするにはサンプルサイズ・データ量が不足

自閉スペクトラム症（ASD）に関する研究

『日本語日常会話コーパス』に含まれる話者のうち、60名を対象に、次の調査を実施：

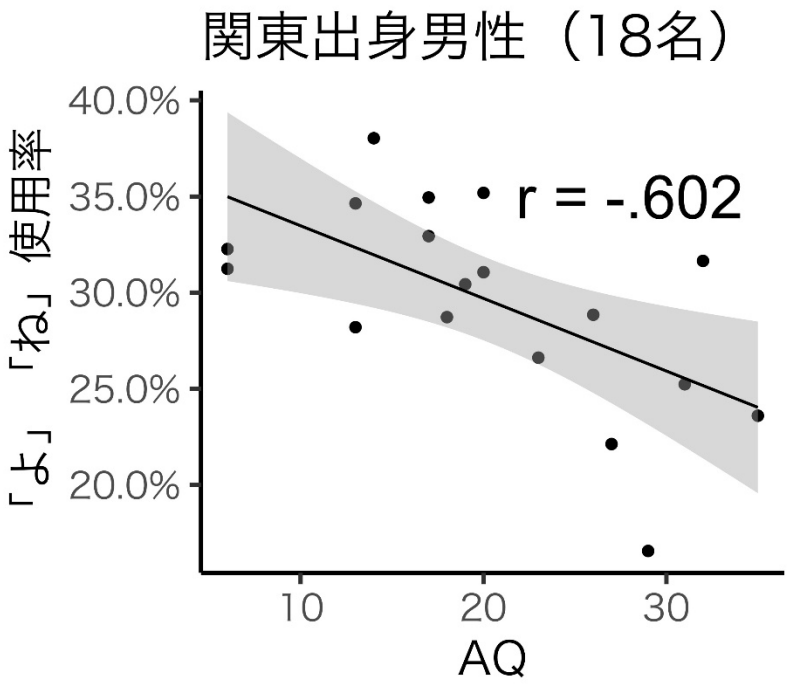
ASDの特性全般に着目

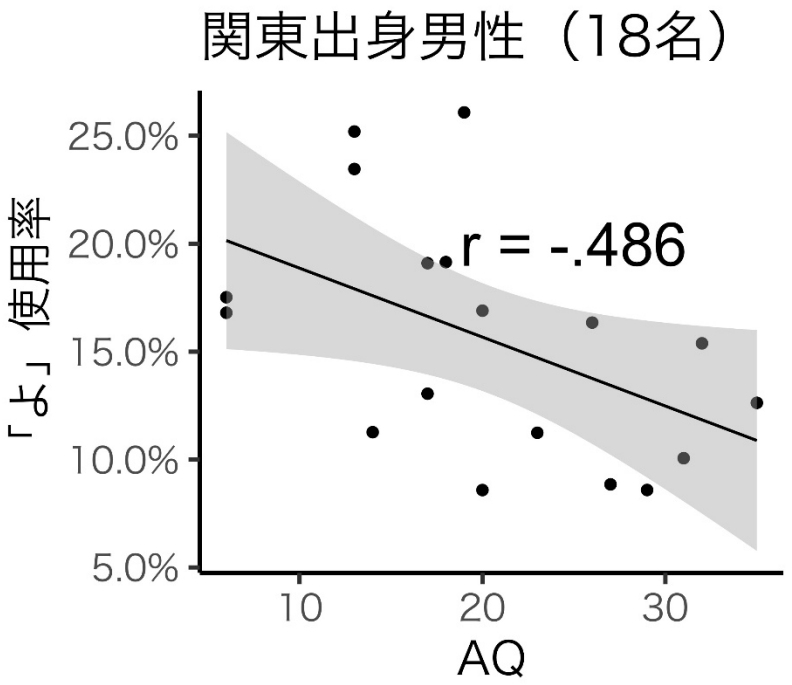
- 自閉症スペクトラム指数 (Autism-Spectrum Quotient , AQ)

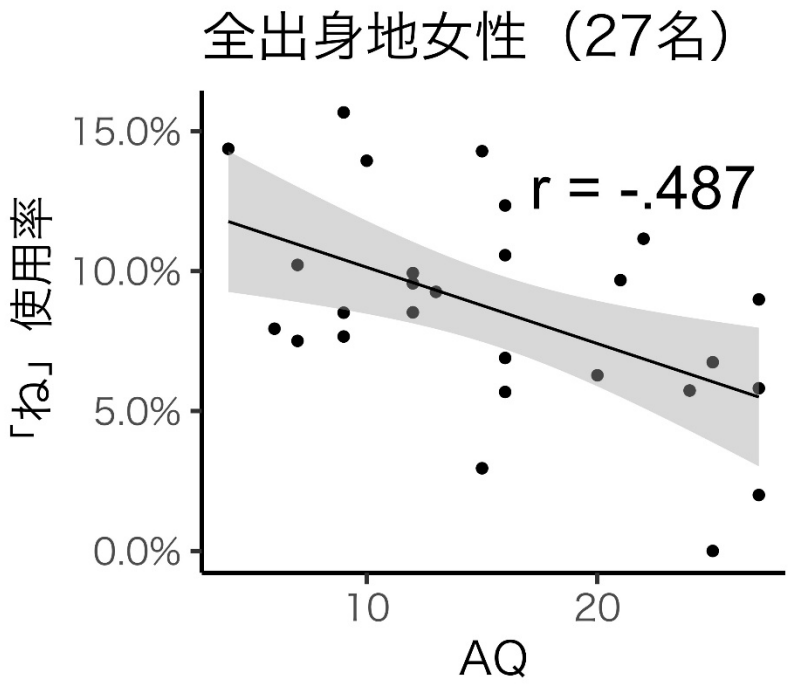
ASDの社会的・情動的特性に着目

- システム化指数 (Systemizing Quotient, SQ)
- 共感化指数 (Empathizing Quotient, EQ)
- 対人反応性指標 (Interpersonal Reactivity Index, IRI)
- トロント・アレキシサイミア尺度 (TAS-20)

自閉スペクトラム症（ASD）に関する研究







自閉傾向が高いほど終助詞使用率が低い

NINJAL
National Institute for Japanese Language and Linguistics

まとめ

- **人文科学分野での利活用**
  - CSJを中心に、音声学を中心に利用
  - CEJCの公開により、談話分析など日本語学・言語学の広い分野で利用
- **情報科学分野での利活用**
  - 音声認識・音声処理・対話処理などで継続的に活用
  - ファインチューニングなど研究課題や技術動向に応じて利用形態は変化
  - 高品質な人手アノテーションデータとして引き続き重要
- **今後の展開**
  - 発達障害、高齢化など社会的課題への応用研究が進展
  - 映像・音声・言語情報を統合したマルチモーダル研究への活用
  - 人文・情報・医療福祉分野を横断する研究基盤としてさらなる発展が期待