UAM corpus tool 分析データの検索と基礎統計処理

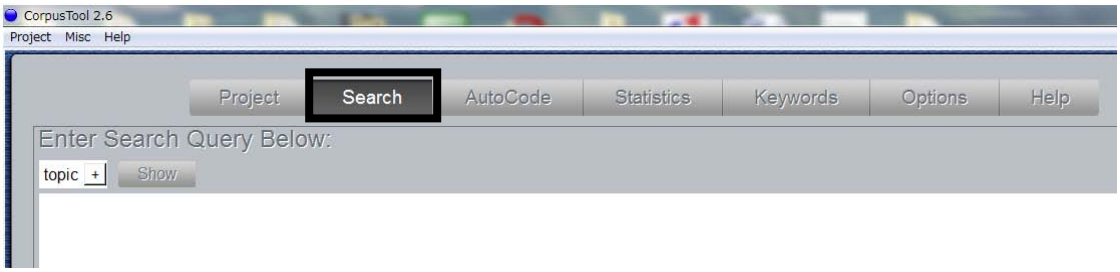
UAM corpus tool の紹介は、今回で最後となります。ここでは、分析データの検索・抽出方法、及び、基本的な統計処理の方法について説明します。なお、以下のものが準備されているという前提で、話を進めていきます。以下のものが準備できていない場合は、第 2 回～第 3 回の資料を見て作成してください。

1. 分析対象となるテクストの ‘incorporate’
2. 分析にもちいるシステムネットワークの記述
3. 少なくとも 2 テクストの分析

① データ（用例）の検索

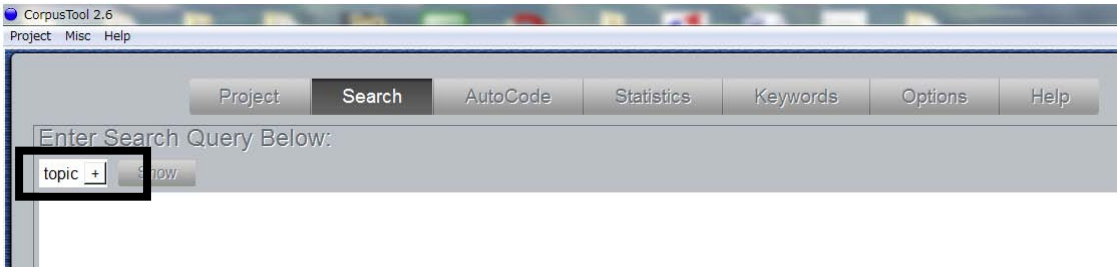
まずは、分析結果を用いてデータを検索する方法について説明します。

【手順 1】メイン画面の上にある「Search」をクリックします。

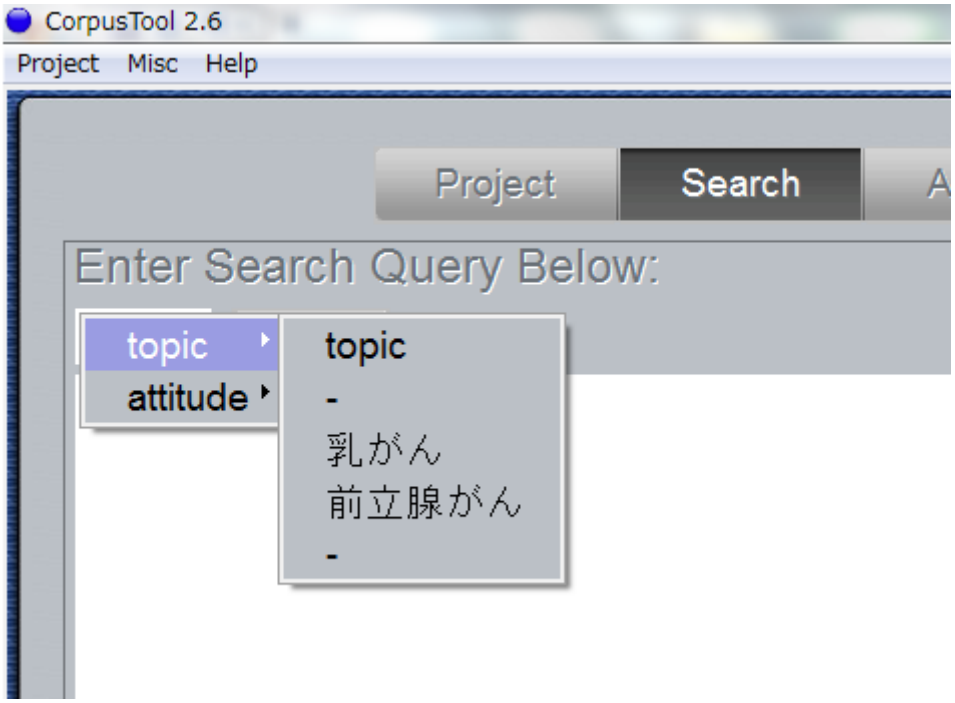


すると、「Enter search Query Below:」と表示されるウィンドウに切り替わります。分析データの検索は基本的にこのウィンドウにて行います。

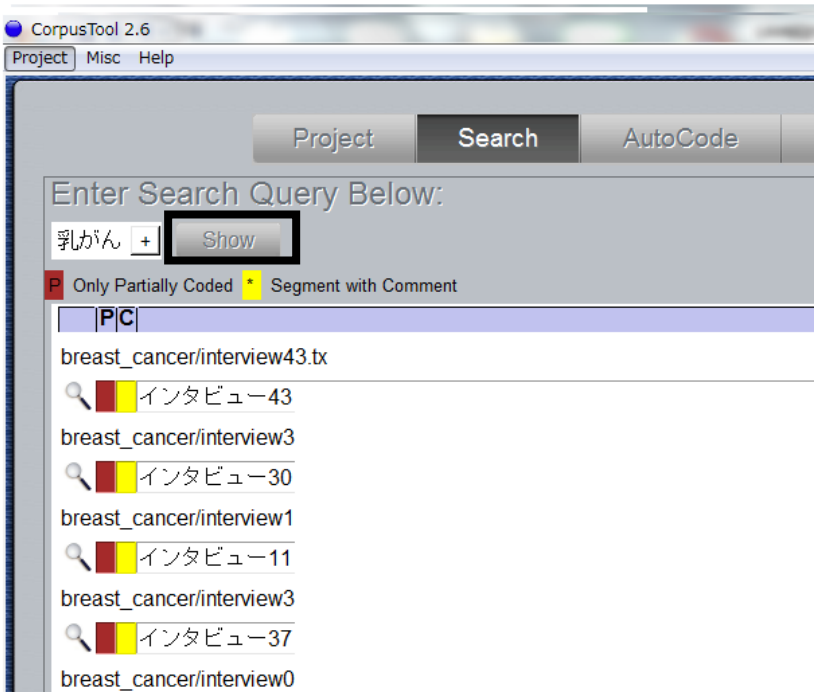
【手順 2】次に、検索したい選択肢(feature)を選択します。まず、画面左のここでは「topic」となっている場所（「＋」ではありません）をクリックしてください。ここでは「topic」となっています（自分が記述したシステムネットワークの point of origin[一番最初の entry condition]が表示されているはずです)。



すると、例えば以下のような画面が表示されます。このリストにシステムネットワークに記述した全選択肢(feature)が表示されるので、リストから選択してください。多層的分析（2 つ以上のシステムネットワークを利用している場合）は、2 つの point of origin が表示されます。

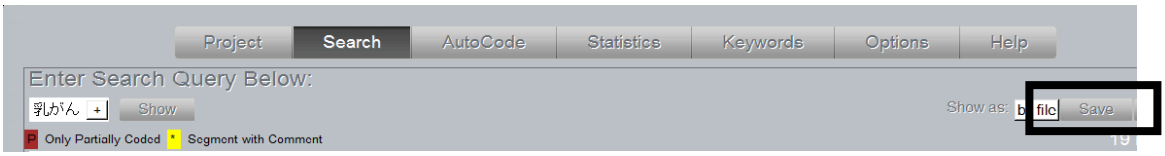


【手順 3】「Show」をクリックします。すると、検索結果が表示されます。

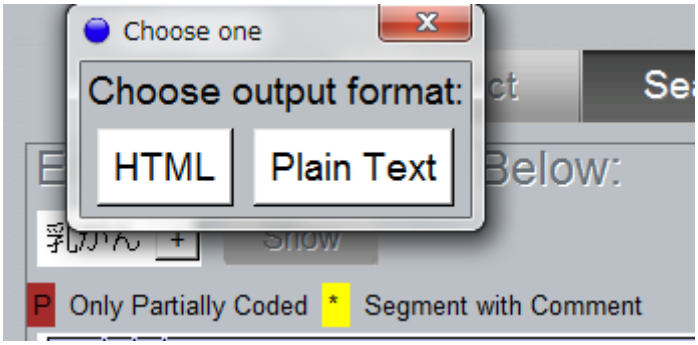


をクリックすると、テクスト全体が表示されます。 に「P」とついているものは、選択肢を全て選びきっていない（分析を途中でやめているもの）です。 に「*」が入っているものは、分析（アノテーション）の際にコメントを残したものです。これが基本的な、分析データの検索方法になります。

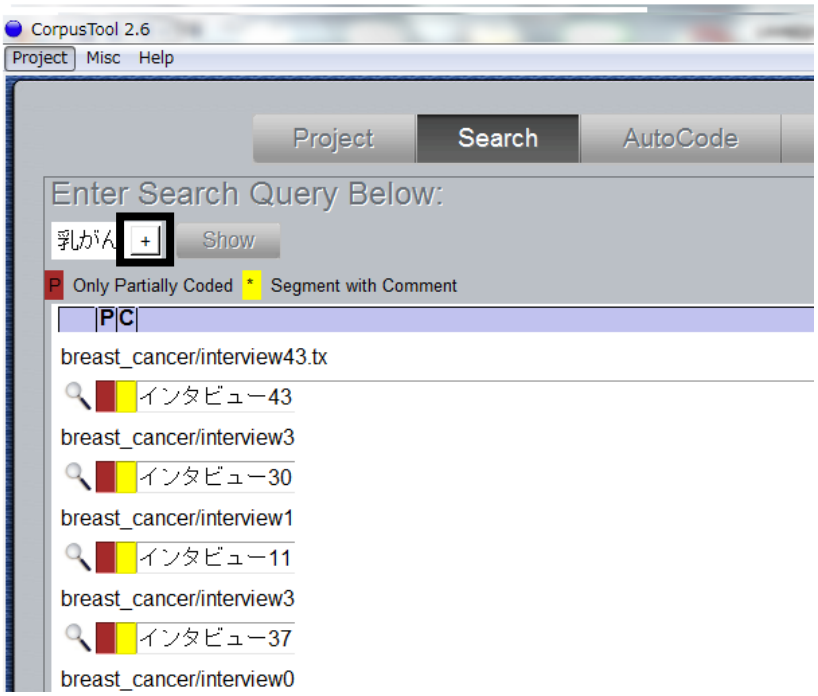
【オプション①】検索結果を保存するには、画面右の「Save」をクリックします。



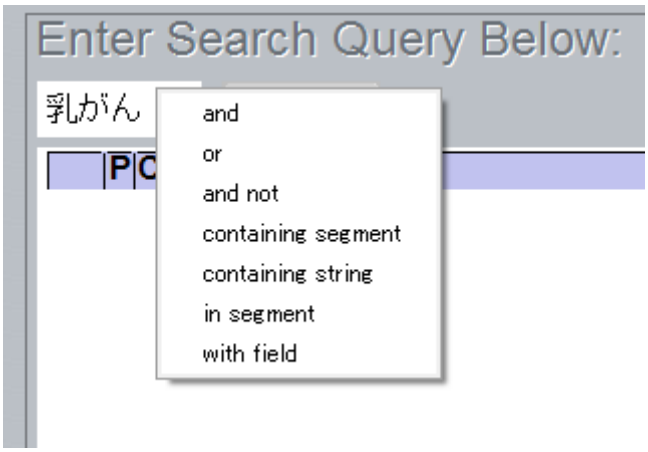
すると、以下のような画面が表示されますので、「HTML」か「Plain Text」のいずれかを選択してください。任意の場所に保存し、HTML はインターネットブラウザで Plain Text はテクストエディターなどで開いてください。



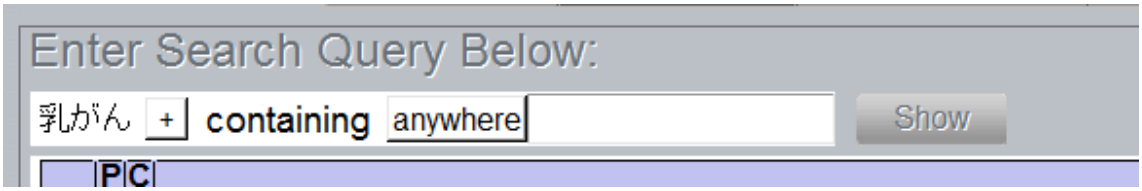
【オプション②】検索条件を追加するには、「＋」をクリックします。



すると、以下のような画面が表示されます。feature を使って条件を絞り込むには（feature A かつ B）「and」を、他の feature も同時に表示するには(feature A もしくは B)「or」を用います。様々な検索ができるので試してみてください。ここでは、「containing string」を使って「乳がん」とアノテーションされているテクストで、かつ、「ショック」という文字列を含むテクストを検索してみます。



まず、「乳がん」のとなりの」「＋」をクリックして、上の画面を表示させ「containing string」を選択します。すると、以下のような画面が表示されます。



次に、「anywhere」のとなりの空欄に「*ショック*」と入力します。「*」はワイルドカード（ショックの前後はどんな文字列でもよい）です。「Show」をクリックして結果を表示させます（著作権の問題で、ここでは文章は表示されていません)。

以上で、分析データの検索方法の説明を終わります。論文を書いていて用例を検索する際や、分析の途中で一度以前分析した言葉がでてきたときに、データを見直す際などに

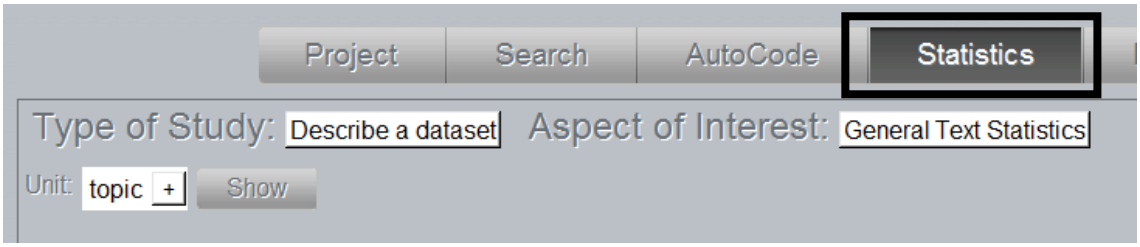
便利です。どのような言葉が特定の feature を具現(realize)するのか、考察する際も便利です。

② 基礎統計処理

ここでは分析データから、基本的な統計情報を表示する方法について説明します。最初にコーパス全体での feature の頻度・プロバビリティの表示方法について説明します。

2-1 コーパス全体での頻度集計・プロバビリティの表示

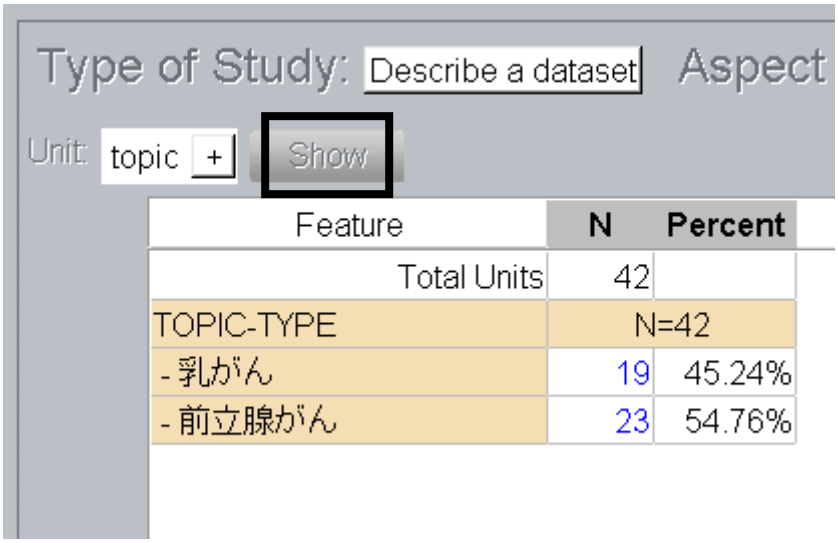
【手順 1】まず、メイン画面から「Statistics」をクリックします。



【手順 2】つぎに、Aspect of Interest を「Feature Coding」に変更します（残念ながら現時点で、「General Text Statistics」 は日本語の分析では機能しません)。

【手順 3】「Unit」で、頻度・プロバビリティを表示したい feature を選択します。ここでは、「topic」が選択されています。

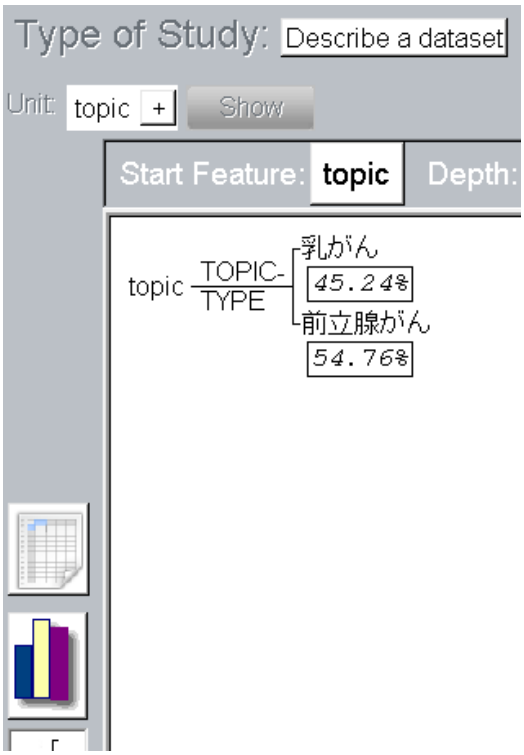
【手順 4】Show をクリックすると、頻度が表示されます。



Feature	N	Percent
Total Units	42	
TOPIC-TYPE	N=42	
- 乳がん	19	45.24%
- 前立腺がん	23	54.76%

＜豆知識＞青い数字で表示されている頻度をクリックすると、当該のデータが「Search」に表示されます。

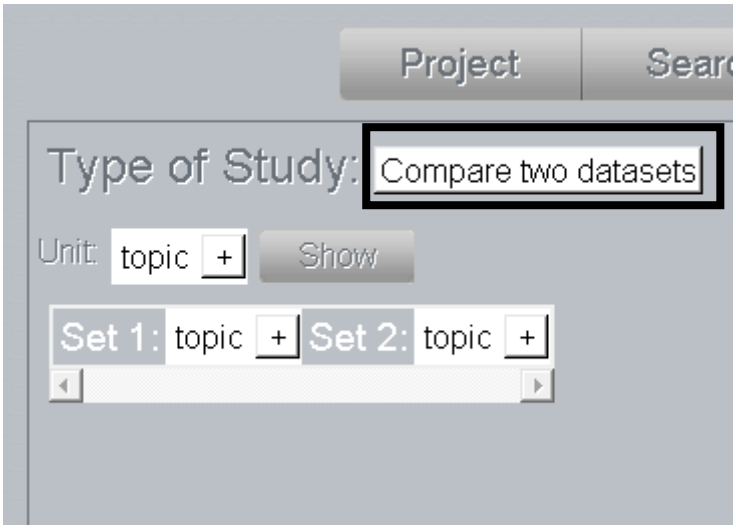
【手順 5】画面左の  をクリックすると、テーブル表示からプロバビリティ表示へ切り替わります。



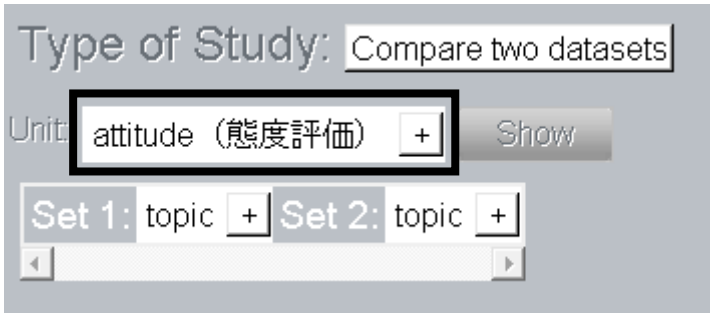
以上で、コーパス全体での頻度の集計・プロバビリティの表示方法の説明を終わります。
＜豆知識＞検索の場合と同様、「＋」をクリックして、頻度を集計したいデータを絞りこむことができます。

2-2 2 つのデータセットの比較

【手順 1】2 つのデータセットにおける頻度の比較を行う際には（例えば、あるジャンルと他のジャンルで頻度を比較する場合など）「Type of Study」を「Compare two datasets」に変更します。

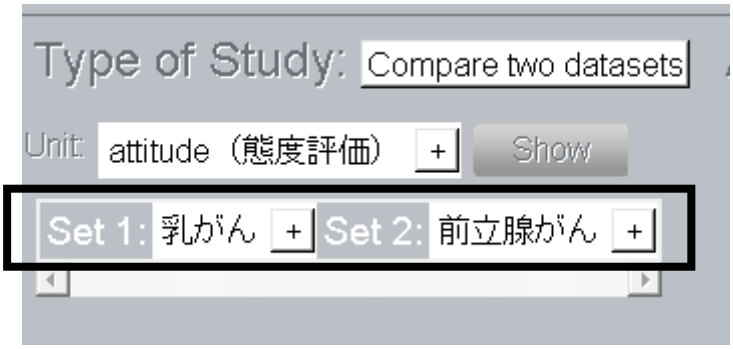


【手順 2】つぎに、頻度を数えたい feature を「Unit」で指定します。ここでは「attitude（態度評価）」となっています。

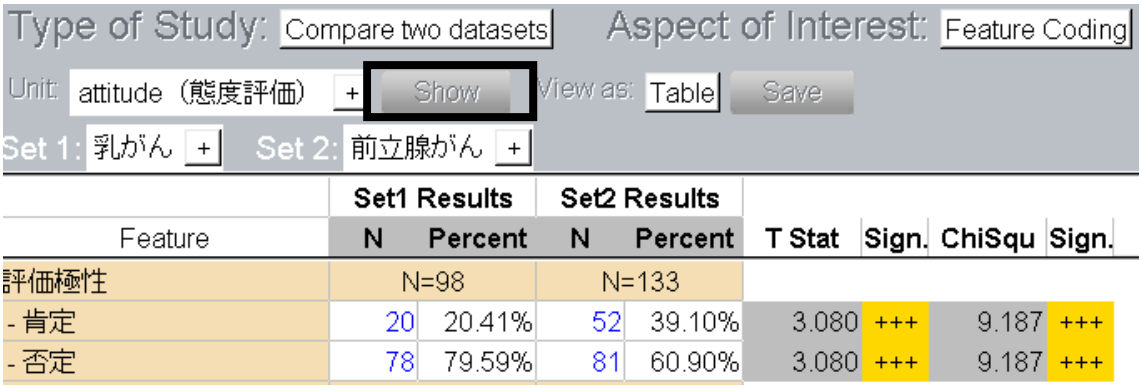


＜豆知識＞ここでも、「＋」を使ってより複雑な条件を設定できます。

【手順 3】比較したいデータセットを選択します。データセットは「Set 1」「Set 2」で選びます。ここでは、トピックが「乳がん」のテクストと「前立腺がん」のテクストとで、「attitude（態度評価）」における feature の頻度の比較を行うよう、設定してあります。



【手順 4】「Show」をクリックします。すると、以下のような画面があらわれます。

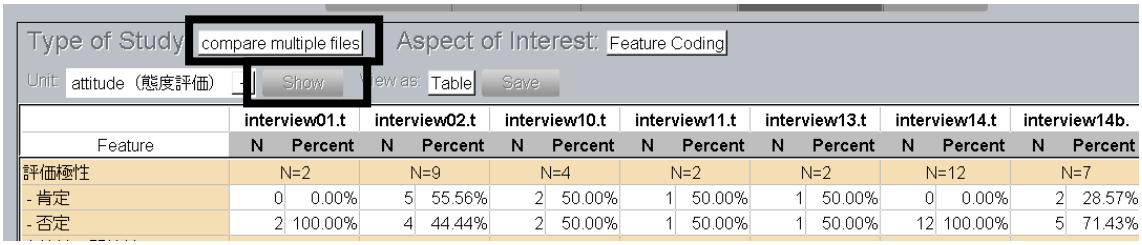


	Set1 Results		Set2 Results					
Feature	N	Percent	N	Percent	T Stat	Sign.	ChiSqu	Sign.
評価極性	N=98		N=133					
- 肯定	20	20.41%	52	39.10%	3.080	+++	9.187	+++
- 否定	78	79.59%	81	60.90%	3.080	+++	9.187	+++

各データセットの頻度及び、t 検定、及び、χ^2 検定の結果が表示されます（但し、適切でないデータに対しても検定をしてしまう場合があるまで、あくまで参考として捉えましょう。論文などに書く際は、ちゃんと R や SPSS などで検定してください）。以上で、2 つのデータセットの比較の説明を終わります。

2-3 各テクストにおける頻度

各テクストごとの頻度を調べたい場合は、Type of Study を「compare multiple files」に設定し、「Show」をクリックします。



	interview01.t		interview02.t		interview10.t		interview11.t		interview13.t		interview14.t		interview14b.	
Feature	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
評価極性	N=2		N=9		N=4		N=2		N=2		N=12		N=7	
- 肯定	0	0.00%	5	55.56%	2	50.00%	1	50.00%	1	50.00%	0	0.00%	2	28.57%
- 否定	2	100.00%	4	44.44%	2	50.00%	1	50.00%	1	50.00%	12	100.00%	5	71.43%

以上で UAM corpus tool の基本的な使用方法の説明を終わりにします。

なお、あくまで裏技としてですが、多量のテクストを扱う際は、ウィザードを使用せずテクストを incorporate する方法があります。また、プロジェクトが収められているフォルダのなかの「Analysis」フォルダには、分析結果が xml ファイルとして保存してあり、この xml ファイルを excel で取り込んだり、自分で XSLT を組めば、様々な形式で結果を表示することができます。この xml ファイルがあれば、分析したテクストを再配布せずに、アノテーション結果のみを公開することができます。色々試してみてください。

****

以上