過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査

村田 真樹 *1　一井 康二 *2　馬 青 *3,*1　白土 保 *1　井佐原 均 *1

*1 情報通信研究機構　*2 港湾空港技術研究所　*3 龍谷大学　(murata @nict.go.jp)

1　はじめに

言語処理学会も発足より 10 年以上経った．本研究では，第 1 回から第 10 回までの 10 年間において，言語処理学会論文誌・年次大会発表においてどういった研究がなされてきたかを調べた．調査方法は，言語処理学会論文誌・年次大会発表の書誌情報を電子化し，その電子的なデータに対して言語処理技術を利用して，種々の有用な研究動向に関する情報を抽出した．この研究では，各研究機関の発表件数の推移や，各研究分野の発表件数の推移も調査した[1]．本研究は言語処理学会における研究動向の調査に役に立つ．動向調査を行なった先行文献としては文献 [2, 3] などがある．

2　研究動向調査

論文誌・年次大会での発表件数の推移を図 1に示す．論文誌と年次大会を比較すると年次大会の方が圧倒的に発表件数が多いことがわかる．また，両方ともに一時期件数が減る時期 (論文誌だと 8 年次，年次大会だと 6 年次) があったものの，全体的に増加傾向であることがわかる．

年次大会と論文誌はそれぞれ 4 年次，6 年次で一時的なピークを迎え，6 年次，8 年次で一時的に落ち込んでいる．年次大会のピークと落ち込みの起こった年の丁度 2 年後に論文誌でピークと落ち込みが生じている．これは，論文誌は投稿から査読，掲載と時間がかかるため，同時期に行なった研究であっても論文誌の方が年次大会よりも遅れて発表されるために生じたものと思われる．(正確には年次大会は論文誌が発行された 1 月から 12 月の次の年の 3 月に行われるため，年次大会に対する論文誌の時間的な遅れは 2 年よりは少し短いものと思われる．)

次に各研究機関の発表件数の推移を調べてみた．それを図 2と図 3に示す．この図では文献 [4] を参考にして等高線表示を利用した．等高線の高さ (色の濃さ) が件数を示す．各研究機関の発表件数のデータにおいて，発表される年次の平均値と最頻値と中央値を求めてそれらの平均を求め，図ではこの平均の値の小さい順に表示した．各研究機関には合計件数と上述の平均の値を付記した．このため，図では早い年次に偏って発表件数の多い機関は上の方に，遅い年次に偏って発表件数の多い機関は下の方に表示される．ここでは合計件数の多かった組織のみを表示した．機関名が変わった組織については頻度の最も大きかった名称を利用して表示している．

これらの図から論文では ATR，通信総研が発表件数が多く，年次大会では NTT， ATR，東工大，通信総研，東京大学の発表件数が多いことがわかる．また， NTT， ATR は早い年次から多くの発表をしているが，通信総研と東京大学は 10 年の年次の中では比較的後ろの方の年次で多くの発表をしていることがわか

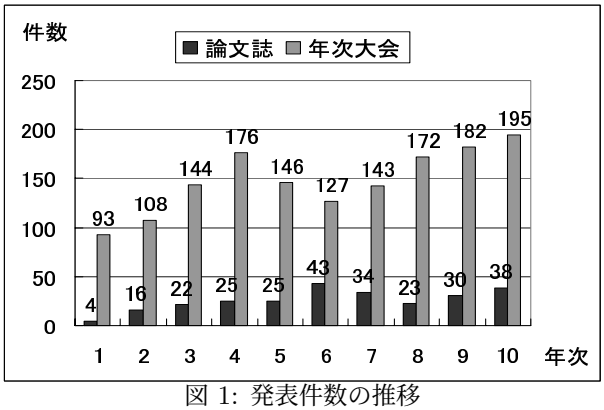

図 1: 発表件数の推移

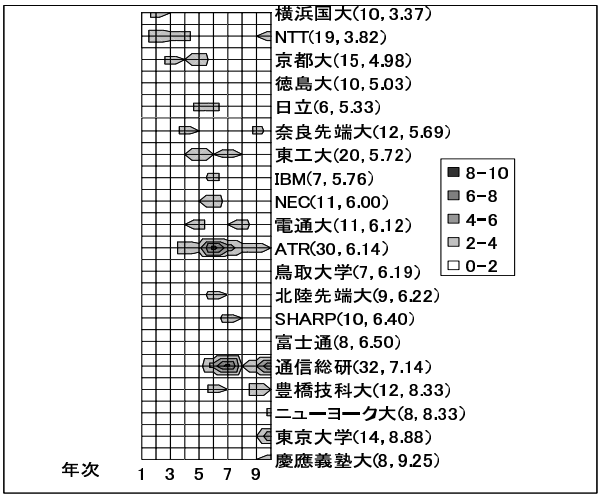

図 2: 研究機関ごとの発表件数の推移 (論文誌)(図の二つの数字は左が合計件数を右が発表件数の多い年次の平均を意味する (厳密な定義は本文を参照のこと))

[1]本稿は文献 [1] において行なった言語処理学会における研究動向調査をより発展させたものである．また本研究は京都大学金丸敏幸氏にお手伝いいただいた．

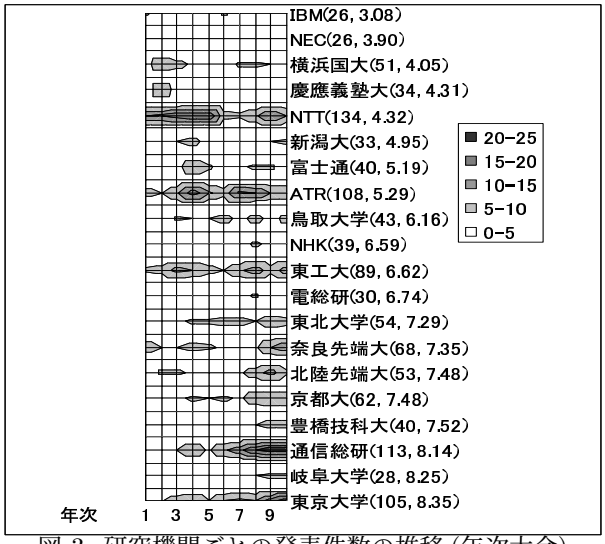

図 3: 研究機関ごとの発表件数の推移 (年次大会)

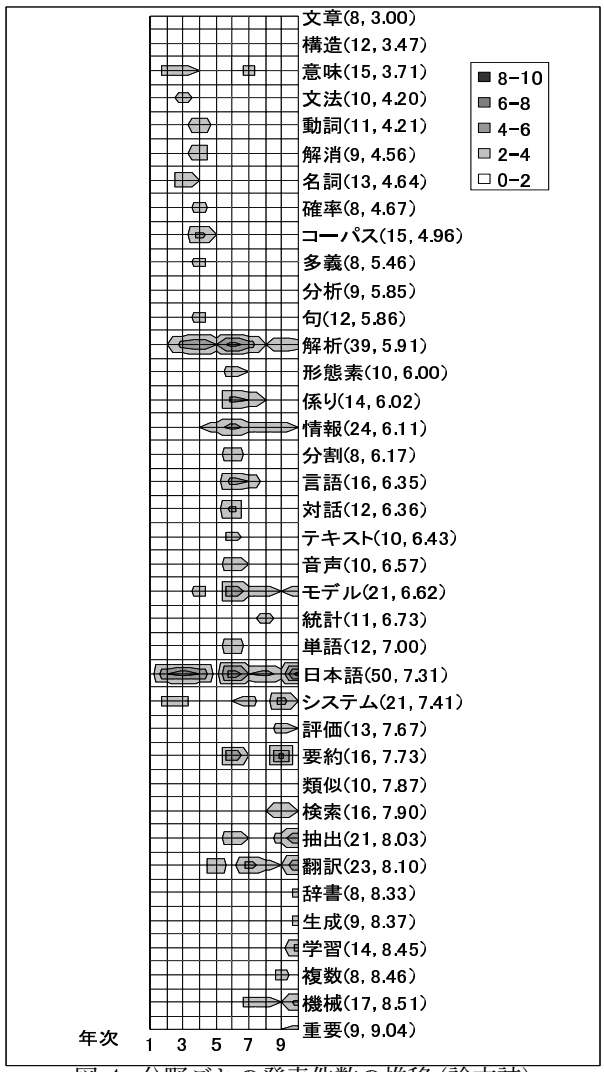

図 4: 分野ごとの発表件数の推移 (論文誌)

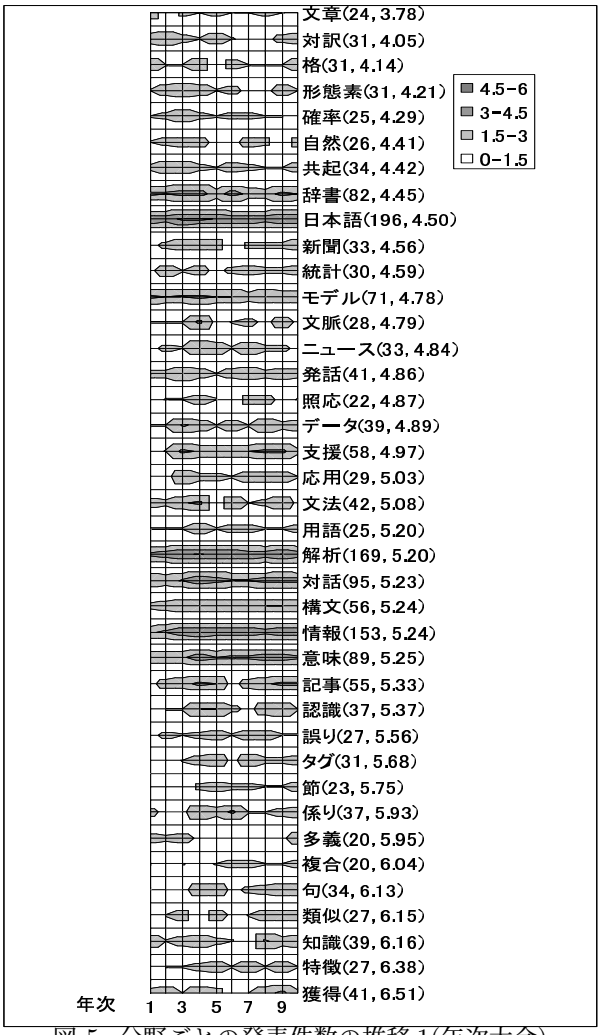

図 5: 分野ごとの発表件数の推移 1(年次大会)

る．通信総研と東京大学は図から増加傾向にあることが読み取れ今後も発表件数が増加することが予測される．その他の組織についてもどの年次で多く発表しているかはこの図を参照することで容易に知ることができる．

次に各研究分野ごとの発表件数の推移を調べてみた．それを図 4から図 6に示す (年次大会は量が多かったので二つの図に分けている)．等高線の高さは論文誌では件数を年次大会では件数に 1 を加えたものの底が 2 の対数を示す．この調査でも各分野ごとの発表件数のデータにおいて，発表される年次の平均値と最頻値と中央値を求めてそれらの平均を求め，図ではこの平均の値の小さい順に表示した．各研究分野には合計件数と上述の平均の値を付記した．ここでは便宜的に Chasen[5] でタイトルを分割することによって得られる各形態素を研究分野とした．タイトルがその形態素を含む論文をその形態素の分野の論文と扱った．研究分野としてふさわしくないもの (例:「的」「研究」)

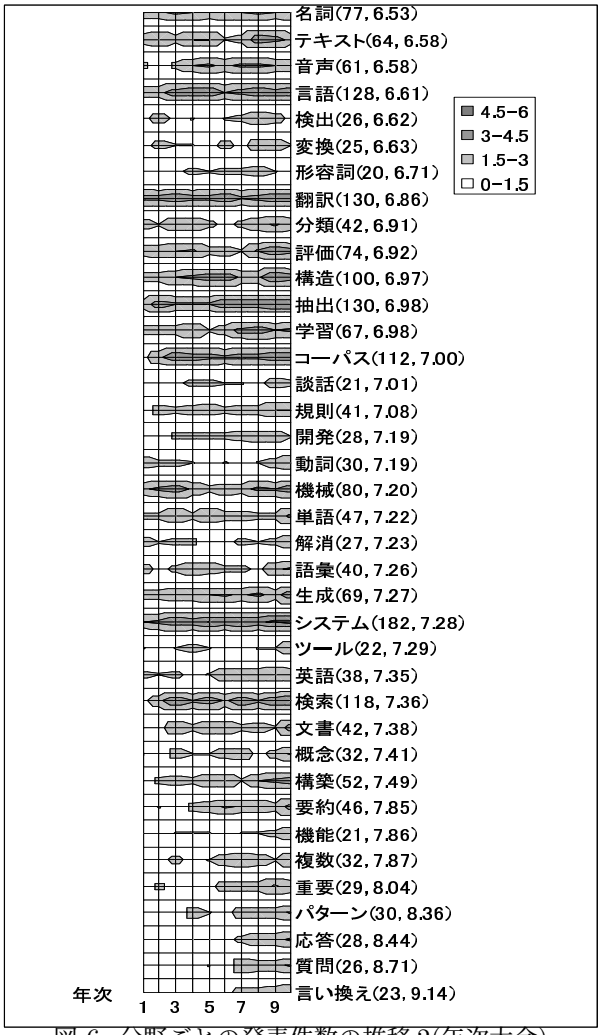

図 6: 分野ごとの発表件数の推移 2(年次大会)

は人手で取り除いた.

図から「日本語」「解析」が特に多いことがわかる. 論文誌では,「動詞」「名詞」「解消」「確率」「コーパス」「多義」などが図の上の方に現れ, これらの研究分野が早い年次に盛んであったことがわかる.「形態素」「係り」「対話」「音声」は 6 年次に盛んであること,「要約」「検索」「翻訳」などが遅い年次で盛んになったこともわかる. 特に「要約」は 6 年次と 9 年次でそれの特集号が出たためそのときに偏って多く出現している.「翻訳」は増加傾向にあることがうかがえ, 今後も発表件数が増加することが予測される.

年次大会では「対訳」「形態素」「確率」「辞書」「統計」などが早い年次で盛んであったことがわかる. また, 下の方を見ると「検索」「要約」「質問」「言い換え」などがあり, これらが最近盛んになっていることがわかる. その他の研究分野についてもどの年次で多く研究されているかはこの図を参照することで容易に知ることができる.

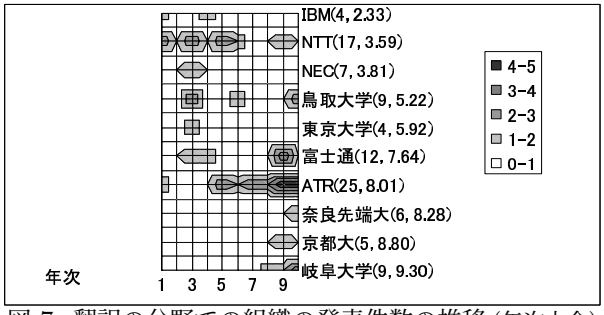

図 7: 翻訳の分野での組織の発表件数の推移 (年次大会)

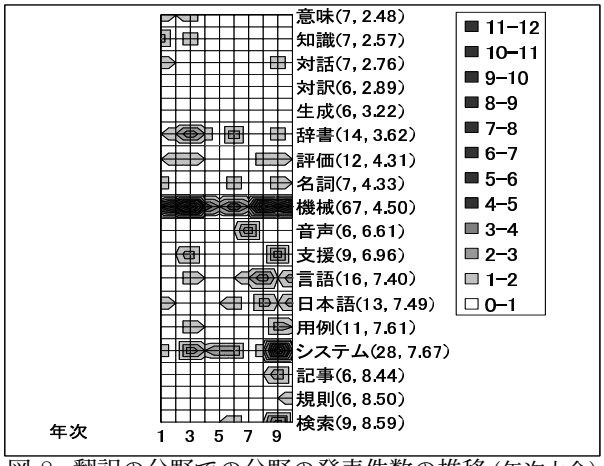

図 8: 翻訳の分野での分野の発表件数の推移 (年次大会)

ここまでは全データを使って調査を行なったが, データの一部を使って同じ調査をしてもよい. 例として翻訳の研究動向を詳しく調べることとし, 年次大会のデータで翻訳をタイトルに含むデータのみを使って同様の調査を行なった. その結果を図 7と図 8に示す. 図の標高線の高さは発表件数を意味する. 図 7から NTT は過去に ATR は最近翻訳の研究が多いことがわかる. また, 図 8から翻訳の研究は過去は「意味」「知識」「辞書」を扱ったものが多かったが, 最近は「支援」「用例」「検索」を扱ったものが多いことがわかる.

最後に 10 年次までのデータで, どの組織がどの研究分野を多く研究しているかを調べてみた. ここでは論文誌における調査結果のみを示す. 組織と研究分野の抽出方法は上述と同じ方法を用いた. 組織と研究分野の共起頻度を求めそのデータに対して双対尺度法 [6] を適用して組織と研究分野の関連性を求めた. それを図 9に示す.

図 9では左下に「翻訳」, 右下には「学習」, 右上には「確率」「検索」, 左上には「名詞」「構造」が現れており, その近辺にそれらに関連する研究分野, 研究組織が現れている. 例えば, 右上の「確率」「検索」の近くにはその研究をよく行なっている日立, 徳島大が現れている. その他にも原点より少し右上のと

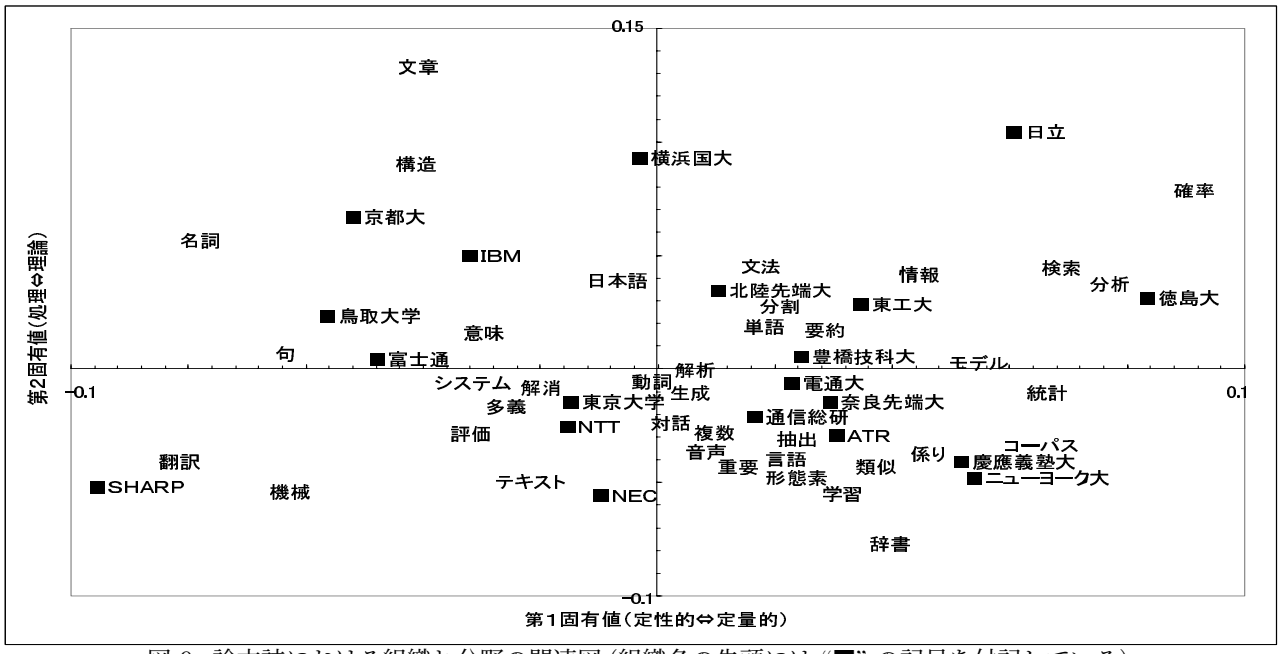

図 9: 論文誌における組織と分野の関連図 (組織名の先頭には “■” の記号を付記している)

ころには「要約」があり，要約の研究を頻繁に行なっている北陸先端大，豊橋技科大，東工大がこれを囲む形で位置している．図を利用することで各研究組織がどのような研究を主に取り組んでいるかを容易に把握することができる．図 9は右に数値的な「確率」「学習」が偏っているため，左に定性的なもの右に定量的なものが配置されているように解釈できる．また，下部に「学習」「翻訳」という複雑な処理のものがあり，上部に「確率」「文法」「文章」「名詞」という理論的なものがあるため，上に理論的なもの下に処理的なものが配置されているように解釈できる．

以上，種々の調査を行なった．本稿では記述しないが，年次と所属または分野の調査においても双対尺度法の分析をしてもよく，所属と分野の調査においても等高線表示を利用してもよい．また，本稿の等高線表示では所属または分野を出現の多い年次の平均の値の順で並べたが，合計発表件数の順で並べたり，所属または分野をその共起語に基づいてクラスタリングしそのクラスタリングの結果類似していると判断されたもの同士が近くに配置されるように並べてもよい．

3 おわりに

本稿では第 1 回から第 10 回までの 10 年間における言語処理学会論文誌・年次大会発表の研究動向の調査を行なった．研究分野については Chasen により分割された形態素を研究分野を示す指標として利用することにより，研究分野に関する研究動向も簡便に調査することができた．調査結果では，論文誌・年次大会における各研究機関や各研究分野の時系列的推移を可視的に示した．また，双対尺度法により研究機関と研究分野の関連性も可視的に示した．本稿の手法は簡便であり，この簡便な手法で多くの示唆を生む結果を示すことができた．

今後は本稿の手法を，情報処理学会，人工知能学会や，さらに多くの種類の学会 (例：土木学会) の書誌情報のデータに適用していきたい．また，研究動向調査に限らず，文献 [7] の問題を含め，社会動向などあらゆる動向調査の問題にも適用していきたい．

参考文献

[1] 村田真樹, 自然言語処理 Vol.11 No.3 編集後記, 言語処理学会誌, Vol. 11, No. 3, (2004.7).

[2] 一條潤子, 石田東生, 谷口守, 黒川洸, 建白白書にみる社会資本整備の歴史的変遷 — キーワード を用いた分析 —, 土木学会第 49 回年次学術講演会 IV-209, (1994), pp. 402–403.

[3] 吉本敏子, 東珠実, 渥見美春, 古寺浩, 鈴木真由子, 菅原亜子, 村尾勇之, アメリカ家政学の系譜 — 学会誌分析 —, 日本家政学会大会 1Ba-4, (1996), p. 70.

[4] 谷口敏夫, 『人工知能と人間／長尾真』のテキスト可視化—KTシステムによるテキスト分析—, (http://www.koka.ac.jp/taniguti96M/0/30/2000/Note2Nagao/Note20000409.htm, 2000).

[5] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, and Masayuki Asahara, Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition, (1999).

[6] 上田太一郎, 刈田正雄, 本田和恵, 実践ワークショップ Excel 徹底活用多変量解析, (秀和システム, 2003).

[7] 加藤恒昭, 松下光範, 平尾努, 動向情報の要約と可視化に関するワークショップの提案, 情報処理学会自然言語処理研究会 2004-NL-164, (2004.11), pp. 89–94.