

FWDUG 発表資料を用いた FWD 英用語集の自動作成の試み

正会員 上島 壯

1. 目的

舗装に錘を落とし、そのたわみを解析するという FWD 試験機が我国に導入されてほぼ 30 年経つ。当初意図された舗装の PMS のための基幹的な装置としての役割とはやや異なる利用のされ方をしているようであるが、舗装体の診断には今も中心的な装置となっている。ところで、海外での FWD の動向について日本の研究者実務者との交流が少なく情報が身近ではない。一方、米国で FWD ユーザーグループ(FWDUG)の発表会が 1992 年に始まり、2000 年以降の発表資料がホームページ上に公開されている。これを利用して今世紀に入ってから FWD の動向をコンピュータを用いて分析できないかという発想から、第一段階として FWD 英用語集の作成を意図し、これをコンピュータで行うことを試みた。

表 -1 FWDUG の発表資料の内容

2. テキストデータベースの作成

FWDUG の発表資料は現在では 2017 年発表まで公開されているが、解析対象はダウンロード時点で最新であった 2015 年までである。発表は表 1 に示すよう

に PPT スライドを主体として、いろいろな様式を含んでいる。その多くを占める PDF, PPT 文書からまずテキストを抽出し、テキストデータベースの構築を行なった。テキスト抽出にはフリーソフト(xdoc2txt)を用いた。多分ソフトの特性でなくテキスト抽出に普遍的に起こることと思われるが、いくつかの問題が生じ、データのクリーニングに多くの時間を要した。たとえば、単語の区切りが正しくなく、並びの単語が細かく切れたり、隙間なく連結する場合がある。また、太字指定の文字列は 2 重出力される。図表、数式を含むページの不規則性はさらに大きい。なお、画像化されている資料については手入力を主体とした。データの管理には自作のソフト(SIRYO2010)を用いたが、その機能として、テキスト内に次の属性行の記号付を行なった。a) タイトル行, b) 著者行, c) 所属機関行, d) 著者連絡情報行, d) 発表出版行, e) 引用文献行, f) キーワード行。この内、今回の用語解析にはタイトル行のみ包含させ、キーワード行については作成作業半ばのため利用しなかった。

3. 単語集の作成

単語の定義：英字，数字，および次の記号 { } - { } { } { } のつながりとする。

文や文節の範囲：ピリオド { . } は基本的に文の終了，また，単なる改行は文の続きと判断し，空行は文の区切りとする，など，いくつかの規則を設定した。

表 2 は，文字列を単語として，英字のみの要素，英字先頭・数字記号後続の要素，その他の要素に区分して集計した結果である。資料全体として 35 万語以上の単語よりなるが，英字のみの単語の種類は 1.1 万語の水準である。個々の単語の表記には異形があり，たとえば，modulus は，その他に MODULUS, MODulus, Modulus の形が出現する。一般に弾性係数の意味であるが，MODULUS という名の FWD 解析ソフトウェアもあって，これの判別は行っていない。用語データの処理は小文字に一元化し，出力される用語集の表記には最も多く現れる形を用い

表 -2 単語作成の段階と結果

要素の区分	延要素数	表記数	一元化表記
英字のみの要素	354,812	16,274	11,264
先頭英字の要素	10,779	3,527	3,059
その他の英数記号	22,079	3,593	3,304
合計	387,670	23,394	17,627

キーワード：FWD, FWDUG, 用語集, 自動作成, テキストデータベース

連絡先：063-0012 札幌市西区福井 1-17-18, Tel 011-663-5652, e-mail khe00340@nifty.com

ている。単語については、別途に用意した一般語辞書と照合して、一般語か非一般語かの属性を付加した。後に用語集として複合語と統合する単語部として、多くの資料に現れる非一般語 470 語と略語 300 語を選出した。選出された単語群には仮の分類コードを付加し、略語には、適宜、スペルアウトした用語と注釈も付加した。

表-3 複合語作成の段階と結果

処理手順	処理1 全素データ	処理2 選択素データ	処理3 一元化表記	処理4 選択一元化表記
2要素用語	367,480	43,165	13,415	4,629
3要素用語	367,216	23,426	15,275	2,839
4要素用語	366,952	10,875	8,788	965
5要素用語	366,688	4,228	3,681	392

表-4 複合語抽出結果例 (資料数最多)

	資料数	総数	抽出複合語
[1]	6	11	spectral analysis of surface waves
[2]	5	5	traffic testing of Airport pavements
[3]	5	5	measured pavement deflections into layer
[4]	5	5	pavement deflections into layer moduli
[5]	13	19	long term pavement performance
[6]	8	9	thickness of pavement layers
[7]	62	117	Falling Weight Deflectometer
[8]	24	32	Portland Cement concrete
[9]	88	276	FWD data
[10]	83	345	data collection

4. 複合語の抽出

複合語は、空白で区切られた複数の要素よりなる文字列で、両端の単語は非一般語であるとした。要素数は最長 5 語の複合語抽出プログラムを作成した。抽出された複合語について表記を一元化してカウントし、1 件のみのものは排除した。また、たとえば、4 要素語には 5 要素語の先頭語か最終語を取り去ったものも含まれる。先頭語を取り去ったと思われる 4 要素語は件数が一致する場合には削除した。

表 3 に処理手順と残留する用語件数を示す。この複合語群と先に選出した単語群を統合して、用語リストを作成し、目で見て吟味する。現在までの吟味作業で 300-400 語の単語の一般語指定を外したが、抽出用語群は顕著に増えた。一般語指定を解除した単語には、east, center, paper, before などが含まれる。

表-4 に抽出複合語の例を、各要素数について出現資料数の多い用語で示す。[1]は SASW と、[5]は LTPP と、通常、略語で用いられる。[3][4]は同じ 6 要素文字列から 5 要素ずつ抽出されたと思われる。抽出した用語群を吟味し、不要語を削除し、シソーラス構築を目標に分類コードを付加するために、表-5 に例示する作業リストを作成した。「分類コード欄」は先に述べた追加単語群データに付加されたものを複合語分統合したものをデフォルトとした。「用語の出現頻度情報」は抽出語の取捨選択の参考にする。

5. おわりに

コンピュータによる文字列パターン照合による共通語探索により、多くの FWD 関係用語をリストアップすることができた。その結果、FWD 関係用語の多くは複合語で構成され、略語化もかなり進んでいったことが分かった。現在の 1 万語レベルの抽出語を整理して、用語検索に必要な十分なデータとして半数程度まで絞り込みたい。

この研究は NPO 法人舗装診断研究会の活動の一環として行なった。データの処理には、北海道科学大学亀山研究室の SAS ソフトウェアを利用させていただいた。ここに謝意を表します。

表-5 複合語選別作業リスト例

分類 コード欄	抽出用語	ペア接続 頻度情報	用語の出現 頻度情報
a40	a) 一般語の設定ミス Don't just collect originally constructed	.3.3... .2...	3 = 2 =
e3	b) 部分要素列の共存 Falling Weight Falling Weight Deflectometer Weight Deflectometer	.*... .*... .*...	145 ===== 117 ===== 135 =====
f3 a30 f0	c) ハイフン接続の使われ方 back calculation back-calculation of layer moduli	.*... .5.*.*.	5 = 3 =
f3 f3 a30 f0 f3 a30 f3 f3 f0	d) 合成語の使われ方 back calculated backcalculated layer moduli backcalculated layer modulus backcalculated moduli	.4... .*... .*... .*...	3 = 13 ===== 2 = 62 =====
A17 a17	e) 略語の説明の付加 ICAS=Inventory Condition and Assessment System /* project ICAS project	.5...	5 =
c2 a20 a22 a20 b1	f) 記号列の用語としての位置づけ P-501 concrete pavement P-501 PCC pavement 18-kip single axle load	.4.*... .3.*... .2.*.*.	4 = 3 = 2 =