한국항해항만학회지 제41권 제3호
J. Navig. Port Res. Vol. 41, No. 3 : 109-118, June 2017 (ISSN:1598-5725(Print)/ISSN:2093-8470(Online))
DOI : http://dx.doi.org/10.5394/KINPR.2017.41.3.109

해양사고 원인을 분류하기 위한 공통단어의 축소에 관한 연구

† 임정빈

† 목포해양대학교 항해학부 교수

A Study on the Reduction of Common Words to Classify Causes of Marine Accidents

† *Jeong-Bin Yim*

† *Division of Navigation Sciences, Mokpo Maritime University, Mokpo 58628, Rep. of Korea*

요 약 : 주제어(key word, KW)는 해양사고의 주요한 원인을 간단하게 표현하기 위한 단어들의 집합으로 해양안전심판원의 심판관들이 작성한다. KW는 심판관들의 서로 다른 주관적인 견해 때문에 일관성 유지가 어렵고, KW의 수가 너무 많은 문제점이 있다. 이러한 문제를 해결하기 위해서는 최적화된 최소의 공통단어(common word, CW)를 이용한 체계적인 KW 구축 프레임이 필요하다. 본 연구의 목적은 체계적인 KW 구축 프레임 개발에 필요한 CW을 도출하는데 있다. 이러한 목적을 달성하기 위하여 본 연구에서는 파레토(Pareto) 분포함수와 파레토 지수를 이용한 최적의 최소 CW 도출방법을 제안하였다. 총 2,642개의 KW을 수집한 후, 수집한 KW의 세부 단어와 이들의 빈도를 갖는 데이터 세트에서 총 56개의 특징적인 CW를 식별하였다. 56개의 특징적인 CW를 이용한 단어 축소실험을 통해서 평균 58.5%의 축소율을 획득하였고, 축소율에 따라서 추정한 CW는 파레토 차트로 검증하였다. 이를 통해서 체계적인 KW 구축 프레임 개발이 가능할 것으로 기대된다.

핵심용어 : 해양사고, 주제어, 원인분류, 단어 축소, 파레토 분포함수, 파레토 지수

Abstract : *The key word (KW) is a set of words to clearly express the important causations of marine accidents; they are determined by a judge in a Korean maritime safety tribunal. The selection of KW currently has two main issues: one is maintaining consistency due to the different subjective opinion of each judge, and the second is the large number of KW currently in use. To overcome the issues, the systematic framework used to construct KW's needs to be optimized with a minimal number of KW's being derived from a set of Common Words (CW). The purpose of this study is to identify a set of CW to develop the systematic KW construction frame. To fulfill the purpose, the word reduction method to find minimum number of CW is proposed using Pareto distribution function and Pareto index. A total of 2,642 KW were compiled and 56 baseline CW were identified in the data sets. These CW, along with their frequency of use across all KW, are reported. Through the word reduction experiments , an average reduction rate of 58.5% was obtained. The estimated CW according to the reduction rates was verified using the Pareto chart. Through this analysis, the development of a systematic KW construction frame is expected to be possible.*

Key words : *Marine Accidents, Key Word, Causation Classification, Word Reduction, Pareto Distribution, Pareto Index*

1. 서 론

해양안전심판원(이하, 해심)에서는 IMO Res. MSC.255(84)(IMO, 1997)에 근거한 「해양사고의 조사 및 심판에 관한 법률 제11690호)」(MOF, 2013)에 의거하여 해양사고를 조사 및 분석하고 있다. 조사 및 분석 결과는 다양한 해양사고 통계 자료(KMST, 2014)와 재결서 및 재결요약서 그리고 해양사고 종류별 주제어 등의 형태로 해양사고조사심판정보포털(이하, 정보포털)(KMST, 2015)에 제공하고 있다.

한편, 해심에서는 해양사고를 상세하게 분석하여 재결서를 작성하고, 재결서를 3쪽 이내로 요약하여 재결요약서를 작성하고 있다. 재결요약서를 작성할 때 심판관들은 해양사고의 주요한 원인을 3~4개의 대표적인 문장으로 기록하고 있는데, 이를 주제어(key word)로 칭하고 있다. 이러한 주제어 작성목적은 해양사고의 주요한 원인을 판단하는 기준으로 활용하고, 용이하게 판례를 검색하기 위한 것이다(KMST, 2007). 따라서 주제어는 해양사고 원인분석과 심판관들의 효율적인 판례 검색에 중요한 데이터이다.

한편, 해심에서는 주제어 작성의 통일성을 기하기 위하여 두 차례에 걸쳐서 주제어를 정비한 바 있다. 1차는 2007년에 실시하였는데, 1963년부터 2007년까지의 재결서 분석을 통하여 해양사고별 대표적인 주제어 2,182개를 결정한 바 있다

† Corresponding author : 종신회원, jbyim@mmu.ac.kr 061) 240-7170
(주) 이 논문은 "해양사고 인적오류 예방을 위한 해심 주제어 분석에 관한 고찰"란 제목으로 "2016 해양과학기술협의회 공동학술대회 한국항해항만학회 프로그램북(부산 벡스코, 2016.5.19-20, pp.196-198)"에 발표되었음.

(KMST, 2007). 그리고 본 논문 저자 등에 의해서 2008년부터 2016년까지 10년간의 재결요약서를 통하여 총 1,311개의 주제어를 2차로 정비한 바 있다(KMST, 2016). 한편, 이러한 2차 정비를 통해서, 선정한 주제어 수가 너무 많고, 시간경과에 따른 주제어 갱신 등이 필요하기 때문에 향후 주제어 축소를 위한 프레임(frame) 개발이 필요한 것으로 나타났다.

이러한 주제어 축소 또는 주제어 구축 프레임에 관한 연구의 필요성은 오래전부터 제기되어 왔는데, Yim(2009a)은 상선 운항사고 평가에 소수의 변수를 적용하기 위한 해양사고 데이터 압축을 시도한 바 있다. 이어서 Yim(2009b)은 압축된 사고원인에 대한 변수를 이용하여 선원의 인적과실을 평가한 바 있다. 그리고 Cho 등(2015)은 차원이 압축된 해심 데이터를 인적모델 개발에 적용하기 위한 연구를 시도한 바 있고, Jang 등(2016)은 현재 해심에서 분류한 주제어 수의 축소를 통한 인적오류 모델 개발에 관해서 보고한 바 있다.

이와 같이 주제어 수의 축소에 관한 연구의 필요성은 본 연구 이전부터 제기되어 왔다. 주제어 수의 축소에 관한 연구의 핵심은 체계적인 주제어 구축 프레임의 개발인데, 이를 위해서는 주제어 작성에 사용하는 단어를 통일시켜야 하고 최소한의 단어를 사용해야 한다. 특히, 최소한의 단어는 단어의 수만 최소화시키는 것이 아니라 사고 내용을 충분하게 설명할 수 있어야 한다. 이러한 단어 중에 가장 중요한 것은 해양사고의 원인을 분류하기 위한 단어(이하, 원인분류단어)이다.

본 연구의 목적은 체계적인 주제어 구축 프레임에 가장 중요한 원인분류단어를 최적의 최소 단어로 축소하기 위한 기법을 개발하여 최소의 단어를 도출하기 위한 것이다.

연구 방법은 다음과 같다. 우선, 해양안전심판원의 정보포털에서 아홉 가지 사고종류에 대한 2,642개의 주제어를 획득한 후, 원인분류단어의 식별에 활용하기 위한 56개의 공통단어를 도출하였다. 그리고 파레토(Pareto) 법칙(Atmour et al., 2014; Finkelstein et al., 2006)에 의거한 파레토 분포함수와 파레토 지수를 이용하여 사고종류별 최적의 최소 원인분류단어를 추정하였다.

2. 연구접근 방법

2.1 연구 접근절차

Fig. 1은 연구 접근절차를 나타낸다. Step 1에서는 주제어(key word, KW)를 수집하여 KW의 특징을 분석하고 체계적인 KW 작성을 위한 프레임(frame)을 검토한다. Step 2에서는 수집한 KW를 정리하여 단어 축소 연구에 적용할 통일된 공통단어(common word, CW)를 식별한다. 그리고 Step 3에서는 CW에서 해양사고 원인분류에 적용할 단어를 식별하고, Step 4에서 파레토(Pareto) 분포함수를 이용한 단어 축소 방법을 개발한다. Step 5에서는 원인분류단어의 축소 작업을 실시하고, Step 6에서 최소의 단어 수를 추정한 후, Step 7에서 추정한 최소 단어를 분석하여 최종적으로 Step 8에서 최적의 원인분류단어를 획득한다.

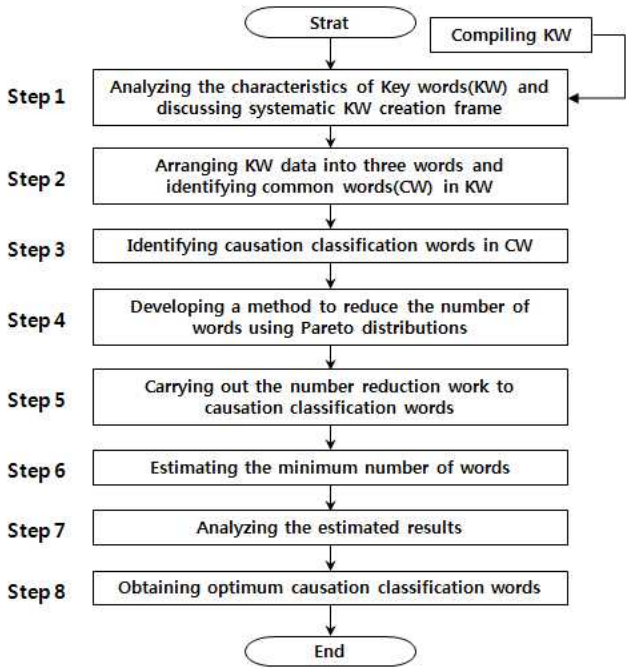

Fig. 1 Study approaching procedures to obtain optimum causation classification words in the establishment of systematic key word frame

2.2 해양사고의 원인분류단어 식별 방법

Fig. 2는 선행연구(KMST, 2016; Jang et al., 2016)에서 제안한 체계적인 KW 구축 프레임을 하나의 예로 나타냈다. 이 프레임은 A부터 F까지의 6개의 변수로 구성되어 있는데, A는 상황을 나타내고, B는 환경, C는 대상, D는 임무, E는 원인, F는 사고종류를 나타낸다. E의 원인은 E-1(인적 원인)과 E-2(물리적 원인)로 세분되고, 변수는 다양한 종속변수를 갖는다.

Fig. 2에 점선의 원과 선으로 표시한 예는, 「항해(A에서 3번) 중, 안개(B에서 3번)가 발생하고 **당직**(C에서 6번) **근무**(D에서 5번) **소홀**(E-1에서 5번)로 충돌(F에서 1번)이 발생한 사고」의 경우를 나타낸 것으로, 그림과 같이 해당하는 단어를 선택하면 자동으로 KW가 당직근무소홀로 작성됨을 나타낸다. 이와 같은 방법을 적용하면, 심판관 모두가 일관된 방법으로 KW을 작성할 수 있을 뿐만 아니라 변수의 조합을 이용하여 통계기반의 해양사고 원인도 분석할 수 있을 것으로 고려된다.

여기서, 원인 변수에 해당하는 종속변수는 해양사고의 원인을 파악하는데 가장 중요한데, 선행연구결과, 종속변수의 종류가 대단히 다양하고 방대한 것으로 나타났다. 따라서 Fig. 2의 프레임을 구축하기 위해서는 원인에 해당하는 종속변수를 소수의 종속변수로 축소하는 것이 우선 필요하다.

그래서 본 연구에서는 기존에 해심에서 사용하고 있는 방

대한 주제어 데이터를 이용하여 E의 원인에 해당하는 최적의 종속변수를 식별하였다.

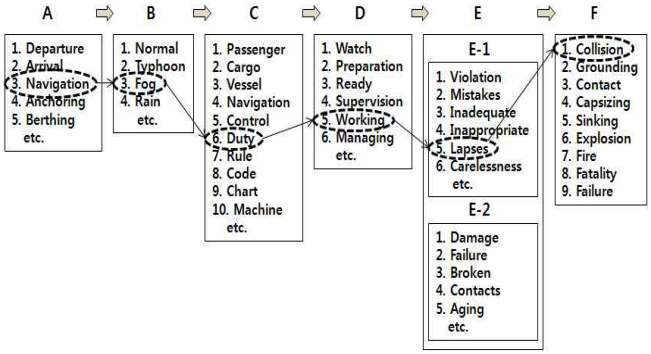

Fig. 2 An example to establish the systematic key word frame

2.3 해양사고의 원인분류 단어 선정 방법

「해양안전심판원 재결판례분석 연구용역 최종보고서(KMST, 2007)」에는 현재 해심에서 사용하고 있는 KW 작성 원칙과 주제어의 특징이 기술되어 있는데, 그 내용을 요약하면 다음과 같다. (a) KW는 재결서 내용을 토대로 작성하고, (b) 한 건의 해양사고에 대해서 3~4개의 KW를 도출한다. (c) 원인판단의 비중이 높은 것을 제1 KW로 선정하고, 그 다음으로 비중이 높은 것을 제2, 제3 등의 KW로 정한다.

이러한 원칙에 의거하여 작성된 KW의 한 가지 예를 들면 다음과 같다. 2003년에 발생한 「어선 제OOO호 기관손상사건」의 경우, 주기관 정비점검소홀, 클러치 손상, 윤활유 공급차단, 기관손상 등 네 가지 KW로 작성되어 있다(KMST, 2003). 이러한 KW는 명사 또는 명사와 형용사의 조합으로 구분할 수 있는 것이 특징인데, 예를 들면 다음과 같다. 주기관 정비점검소홀의 경우는 주기관과 정비점검 그리고 소홀로 구분할 수 있고, 클러치 손상의 경우는 클러치와 손상으로 구분할 수 있다. 이러한 예는 기관손상사건에 관한 것인데, 항해와 관련된 KW의 예를 들면 다음과 같다. 제한시계 경계소홀의 경우는 제한시계와 경계 그리고 소홀로 구분할 수 있고, 협수로 무중항법위반의 경우는 협수로와 무중항법 그리고 위반으로 구분할 수 있다.

위의 예와 같이 구분한 단어 중에서, 첫 번째 단어는 주로 주어진 상황 또는 장비 명칭 등으로 구성되고, 두 번째 단어는 첫 번째 단어에 대한 행위 또는 상황의 적용 등으로 구성되며, 세 번째 단어는 사고의 원인을 나타낸 것으로, 주로 해기사의 인적원인 또는 장비와 기기의 물리적인 원인 등으로 구성되어 있다.

여기서, 첫 번째와 두 번째 단어는 선박에서 일반적으로 사용하는 단어들이 나타나는데 반하여, 세 번째 단어는 심판관들의 주관적인 견해가 반영된 방대한 종류의 단어가 등장한다. 그래서 본 연구에서는 방대한 종류의 세 번째 단어를 원인분류 단어로 선정하였다.

3. 실험 데이터 구축

3.1 주제어 데이터 수집과 분류

해심의 정보포털(KMST, 2015)에서 아홉 가지 사고종류(충돌, 좌초, 접촉, 전복, 화재/폭발, 침몰, 기관고장, 인명사상, 여객사상)에 대한 KW 총 2,982개를 수집한 후, 중복되거나 KW로 부적절한 것 등을 제외하고 총 2,642개의 KW을 선정하였다. 선정한 KW는 액셀파일(Excel file)에 다음과 같은 형태로 분류하였다(Yim et al., 2014).

$\langle A_k, KW_{k,j_k} \rangle$

여기서 〈 〉는 데이터 세트(data sets)를 의미하고, A_k는 사고종류의 단어를 나타낸다. KW_{k,j_k}는 k의 사고종류에 대한 j_k의 KW을 나타내고, $k(k=1,2,3,\cdots,K;K=10)$는 사고종류를 식별하기 위한 인덱스(index)를 나타내며, $j_k(j=1,2,3,\cdots,J_k)$는 k에 따라서 결정되는 KW의 수를 나타낸다.

한편, $KW_{k,j_k} \in \{\omega 1_{k,j_k}, \omega 2_{k,j_k}, \omega 3_{k,j_k}\}$는 3가지 단어로 구성하였는데, 여기서 $\omega 1_{k,j_k}$, $\omega 2_{k,j_k}$, $\omega 3_{k,j_k}$는 위의 2.3절의 예에서 설명한 첫 번째, 두 번째, 세 번째 단어를 각각 나타낸 것으로, $\omega 3_{k,j_k}$에 본 연구대상인 원인분류단어가 포함되어 있다. 예를 들어 위의 2.3절에 나타낸 제한시계 경계소홀의 경우, $\omega 1$은 제한시계, $\omega 2$는 경계, $\omega 3$는 소홀로 구분하는 것이다.

위와 같이 수집하고 분류한 KW의 수를 다음 식(1)으로 계산하여 Table 1에 나타냈다.

$$NKW_k = \sum_{j=1}^{J_k} I_{k,j_k}, \quad \begin{cases} I_{k,j_k}=1 & \text{if} \quad KW_{k,j_k} \neq O \\ I_{k,j_k}=0 & \text{if} \quad others \end{cases} \tag{1}$$

여기서 NKW_{k,j_k}는 KW_{k,j_k}에서 k에 해당하는 KW의 수를 나타내고, O는 영(0) 데이터를 나타낸다.

Table 1에서, $NKW1_k$와 $NKW2_k$은 처음에 수집한 KW와 본 연구에 적용한 KW의 수를 나타낸다. 앞에서 설명한 바와 같이 $k=1$부터 $k=9$까지는 아홉 가지 사고종류를 나타내고, $k=10$는 아홉 가지 사고종류 전체의 합을 나타낸다.

여기서, $KW_{k,j_k} \in \{\omega 1_{k,j_k}, \omega 2_{k,j_k}, \omega 3_{k,j_k}\}$에 분류한 $\omega 3_{k,j_k}$의 수는 $NKW2_k$와 같이 방대하고, $\omega 3_{k,j_k}$에는 동일한 단어가 포함되어 있다. 예를 들면, 소홀이라는 단어는 $\omega 3_{k,j_k}$에 중복하여 나타난다. 단어 축소를 위해서는 전체 사고종류에서 출현 빈도가 높게 나타나는 공통된 단어를 우선 식별해야 한다.

Table 1 Compiled key words according to the type of accidents A_i. $NKW1_k$ and $NKW2_k$ denotes the acquired number and the used number of key words in this work, respectively

k	A_k	$NKW1_k$	$NKW2_k$
1	Collisions	559	470
2	Contacts	207	201
3	Grounding	408	374
4	Capsizing	237	193
5	Fire/Explosion	326	289
6	Sinking	486	416
7	Machinery Failure	532	503
8	Crew Causalities	201	170
9	Passenger Causalities	26	26
10	Sum	2,982	2,642

3.2 공통단어 식별

공통단어(CW)는 다음과 같이 식별하였다. 우선, KW_{k,j_k} 에서 아홉 가지 사고종류 전체에 대한 $\omega 3_{k,j_k}(k=10)$ 을 다음과 같이 $W_t^{\cup}$ ($\cup$는 합집합을 의미)로 둔다.

$$W_t^{\cup} = \omega 3_{k,j_k}(k=10) \tag{2}$$

여기서 $t=1,2,3,\cdots,T$ 이고 $T=J_k(k=10)$ 이며 $J_k(k=10)$ 는 Table 2에서 $NKW2_k(k=10)=2,642$ 이다.

$W_t^{\cup}$ 에 중복되어 나타나는 단어를 식별하여 $W_t^{\cap}$ 로 둔다.

$$\begin{cases} W_t^{\cap} = W_t^{\cup} & \text{if } \left((W_t^{\cup} \bigcap_{t=1}^{T} W_{t-1}^{\cup}) = True \right) \\ W_t^{\cap} = O & \text{if } others \end{cases} \tag{3}$$

여기서 $\cap$는 단어의 교집합을 의미한다.

식(3)의 $W_t^{\cap}$ 을 이용하여 동일한 단어가 나타나는 빈도 $\aleph_t^{\cap}$ 을 다음과 같이 계산한다.

$$\aleph_t^{\cap} = \sum_{t=1}^{T} I_t, \quad \begin{cases} I_t = 1 & \text{if } W_t^{\cap} \neq O \\ I_t = 0 & \text{if } others \end{cases} \tag{4}$$

그리고 식(3)의 $W_t^{\cap}$ 와 식(4)의 $\aleph_t^{\cap}$ 로 구성된 데이터 세트 $\langle W_t^{\cap}, \aleph_t^{\cap} \rangle$ 을 구축하고, 조건 H 을 다음과 같이 정한 후, H 을 만족하는 $\aleph_t^{\cap}$ 의 인덱스 IW_t 을 다음 식(5)으로 식별한다.

- 조건 H : $\aleph_t^{\cap} \geq 10$인 단어 또는 $\aleph_t^{\cap} < 10$인 경우에도 해양사고 원인판단에 중요한 단어

$$\begin{cases} IW_t = 1 & \text{if } (H = true \mid \aleph_t^{\cap}) \ (1 \leq t \leq T) \\ IW_t = 0 & \text{if } others \end{cases} \tag{5}$$

위의 $\langle W_t^{\cap}, \aleph_t^{\cap} \rangle$ 에서 $IW_t = 1$ 일 때의 $\langle \grave{W}_m, \grave{\aleph}_m \rangle$ 을 다음과 같이 분리한다.

$$\langle \grave{W}_m, \grave{\aleph}_m \rangle = \left(\langle W_t^{\cap}, \aleph_t^{\cap} \rangle \mid IW_t = 1 \right) (1 \leq t \leq T) \tag{6}$$

여기서 $m=1,2,3,\cdots,M$ 이고, $M = \sum_{t=1}^{T} IW_t$ 이다.

식(6)의 $\langle \grave{W}_m, \grave{\aleph}_m \rangle$ 에서 $\grave{W}_m$ 가 CW에 해당하는 단어가 된다. 다시 $\grave{W}_m$ 에 대한 $\grave{\aleph}_m$ 을 가장 큰 것부터 작은 순서대로 정렬하여 CW에 대한 단어 CW_m 와 빈도 $\aleph_m$ 의 데이터 세트를 식(7)로 구축하였다.

$$\langle CW_m, \aleph_m \rangle = \langle \grave{W}_m, \grave{\aleph}_m \rangle \mid (\grave{\aleph}_{m-1} \leq \grave{\aleph}_m) \tag{7}$$

Table 2에 식(7)으로 구축한 CW_m 와 $\aleph_m$ 을 나타냈다. 총 56개의 CW가 도출되었는데, $m=1$ 의 소홀의 CW_m 가 $\aleph_m = 287$ 로 가장 크고, $m=56$ 의 폭발의 CW_m 가 $\aleph_m = 7$ 로 가장 작게 나타났다.

Table 2 Identified common words CW_m and their frequencies $\aleph_m$

m	CW_m	$\aleph_m$	m	CW_m	$\aleph_m$
1	Lapses	287	29	Excess	22
2	Violation	207	30	Ship feature	21
3	Inadequate	182	31	Unlawfulness	21
4	Inappropriate	173	32	Fire	20
5	Carelessness	118	33	Over	17
6	Nonfulfillment	96	34	Untie	17
7	Badness	92	35	Not indicate	17
8	Damage	81	36	Not confirm	16
9	Unreasonable	71	37	Ablation	15
10	Failure	69	38	Sticking	14
11	Broken	62	39	Leaking	14
12	Inexperienced	52	40	Crack	13
13	Mistakes	47	41	Loose	13
14	Contacts	44	42	Flow	13
15	Aging	43	43	Short	13
16	Cutting	41	44	Over speed	13
17	Negligence	38	45	Parting	12
18	Weather	35	46	Flood	12
19	Insufficient	34	47	Reckless	12
20	Cut and broken	30	48	Drinking	12
21	Poor	29	49	Unidentified	12
22	Delay	29	50	Defect	11
23	Pole	24	51	Leakage	11
24	Perceive fail	24	52	Approach	11
25	Overheating	23	53	Bursting	9
26	Breakaway	23	54	Overestimate	8
27	Wrapping	22	55	Closed	7
28	Dropping	22	56	Explosion	7

3.3 사고종류별 공통단어 식별

Table 2의 CW를 이용하여 사고종류에 각각에 포함된 공통단어를 식별하면 다음과 같다. 먼저, $\omega 3_{k,j_k}$에서 k에 해당하는 CW의 수 NCW_k을 다음과 같이 계산한다.

$$NCW_k = \sum_{j_k=1}^{J_k} I_{j_k}, \tag{8}$$

$$\begin{cases} I_{j_k} = 1 \text{ if } CCW_m(1 \le m \le M) \equiv w3_{k,j_k}(1 \le k \le K, 1 \le j_k \le J_k) \\ I_{j_k} = 0 \text{ if } others \end{cases}$$

여기서 $M(M=56)$ 은 Table 2에 나타낸 CW의 수를 나타내고, J_k 는 Table 1의 $NKW2_k$와 같다.

다음에는 $\omega 3_{k,j_k}$에 해당하는 CCW_m 의 인덱스 m 을 갖는 ICW_{k,n_k} 을 다음과 같이 식별한다.

$$ICW_{k,n_k} = \tag{9}$$
$$m \mid \left(CW_m(1 \le m \le M) \equiv w3_{k,j_k}(1 \le k \le I, 1 \le j_k \le J_k) \right)$$

여기서 $n_k(n_k = 1,2,3,\cdots,N_k)$이고, $N_k = NCW_k$ 이다.

ICW_{k,n_k}에 해당하는 CW_m 과 $\aleph_m$ 을 다음과 같이 도출하여 사고종류별 공통단어 $\grave{CW}_{k,n_k}$ 과 이에 대한 빈도 $\grave{CF}_{k,n_k}$ 로 둔다.

$$\grave{CW}_{k,n_k} = CW_m \mid (m = ICW_{k,n_k}) \tag{10}$$

$$\grave{CF}_{k,n_k} = \aleph_m \mid (m = ICW_{k,n_k}) \tag{11}$$

그리고 데이터 세트 $\left\langle \grave{CW}_{k,n_k}, \grave{CF}_{k,n_k} \right\rangle$을 구성한 후, 다음 식(12)과 같이 $\grave{CF}_{k,n_k}$ 가 큰 것부터 작은 순서대로 정렬하여 사고종류별 공통단어 CW_{k,n_k} 와 빈도 CF_{k,n_k} 의 데이터 세트를 구축하였다.

$$\left\langle CW_{k,n_k}, CF_{k,n_k} \right\rangle = \left\langle \grave{CW}_{k,n_k}, \grave{CF}_{k,n_k} \right\rangle \mid (\grave{CF}_{k,n_k-1} \le \grave{CF}_{k,n_k}) \tag{12}$$

Appendix A에 식(9)의 ICW_{k,n_k} 과 식(12)의 CF_{k,n_k} 을 나타냈다.

4. 단어 압축 방법

4.1 파레토 분포함수와 지수

파레토(Pareto) 법칙은 「20%(q)의 인구가 80%(p)의 수입을 갖고 있다」는 것으로, $p+q=1$ 규칙으로 불린다(Wikipedia, 2016). 본 연구에서는 20%에 해당 단어가 전체단어의 80%를 설명할 수 있는 최소의 단어 수를 구하기 위하여 파레토 함수를 적용하였다.

파레토 함수는 X가 형식 1의 파레토 분포를 갖는 랜덤(random) 변수인 경우 X가 임의의 수 x 보다 클 때의 확률 $\Pr(X>x)$ 로 정의할 수 있고(Brynjolfsson et al., 2007; Fialova et al., 2004), $\Pr(X>x)$ 에 대한 누적분포함수는 $F(x)=1-\Pr(X>x)$로 정의된다(Rytgaard, 1990; Sousa and Michailidis, 2004; Vilar-Zanon and Lozano-Colomer, 2007). 그리고 파레토 지수(index)는 $\alpha = \log(q)/\log(q/p), (p>q)$ 으로 계산할 수 있다(Wikipedia, 2016).

4.2 단어 축소 절차

다음과 같은 4단계 절차를 적용하여 단어를 축소하였다.

Step 1. 설계조건 설정

단어 축소를 위한 설계조건으로, 「사고종류별로 20%의 단어가 80%의 누적빈도를 갖는다.」고 정한다. 파레토 법칙을 따르는 경우, $q=0.2$, $p=0.8$ 가 되고, 이에 대한 파레토 지수 $\overline{\alpha}=\log(0.2)/\log(0.2/0.8)=1.161$이 된다. 만약, 축소한 단어가 $\overline{\alpha}$ 이내의 값이면 설계조건을 만족하고, 그러하지 않은 경우에는 $\overline{\alpha}$ 이내의 값이 되도록 p 와 q 의 비율을 조정하여 단어를 축소해야 한다.

Step 2. 파레토 분석

공통단어에 대한 빈도 CF_{k,n_k} 의 누적분포 $Fp_k(n_k)$ 을 계산한다.

$$Fp_k(n_k) = \left(\sum_{n_k=1}^{N_k} \sum_{j=1}^{J=n_k} CF_{k,j} \right) / \sum_{n_k=1}^{N_k} CF_{k,n_k} \tag{13}$$

여기서 $Fp_k(n_k)$는 p 을 구하기 위한 것으로, k에 대한 $n_k(n_k=1,2,3,\cdots,N_k; N_k = NCW_k)$의 누적분포가 된다.

다음에는 공통단어의 수 NCW_k에서 축소하려는 단어의 수 Nq_k 을 계산한다.

$$Nq_k = NCW_k \times q \tag{14}$$

여기서 q는 축소하려는 단어의 비율을 의미하는 것으로, 설계조건에서 $q=0.2$이다.

$Fp_k(n_k)$에서 Nq_k 에 해당하는 누적빈도 p_k을 구한 후, q_k와 α_k을 계산한다.

$$\begin{cases} p_k = Fp_k(n_k = Nq_k) \\ q_k = 1 - p_k \\ \alpha_k = \log(q_k)/\log(q_k/p_k), \quad p_k > q_k \end{cases} \tag{15}$$

만약 식(15)의 계산 결과가 $\alpha_k \le \overline{\alpha}$ 이면 식(14)의 Nq_k을 축소하려는 단어의 수로 정하고, $\alpha_k > \overline{\alpha}$ 인 경우에는 다음과 같은 절차에 따라서 Nq_k을 추정한다.

Step 3. $\alpha_k > \overline{\alpha}$ 인 경우의 단어 축소

$Fp_k(n_k)$에서 주어진 조건을 만족할 때의 단어의 수 $\widehat{Nq}_k$을 다음과 같이 추정한다.

$$\widehat{Nq}_k = (\min(n_k) \mid (Fp_k(n_k) \ge p))\,(1 \le n_k \le N_k) \tag{16}$$

여기서 $\widehat{Nq}_i$는 주어진 조건을 만족하는 n_k의 최솟값 식별 결과를 나타낸 것으로, 축소한 단어의 수가 되고, p는 원하는 누적빈도를 의미하는 것으로 설계조건에서 $p = 0.8$ 이다.

다음과 같이 $\widehat{Nq}_k$을 이용하여 $\hat{p}_k$와 $\hat{q}_k$을 계산한 후, $\hat{\alpha}_k$을 이용하여 축소한 단어가 파레토 법칙을 따르는지 평가한다.

$$\begin{cases} \hat{p}_k = Fp_k(n_k = \widehat{Nq}_k) \\ \hat{q}_k = 1 - \hat{p}_k \\ \hat{\alpha}_k = \log(\hat{q}_k)/\log(\hat{q}_k/\hat{p}_k), \quad \hat{p}_k > \hat{q}_k \end{cases} \tag{17}$$

Step 4. 축소율 평가

단어 축소결과는 축소율 $\gamma_k(\%)$로 평가한다.

$$\gamma_k = 1 - \frac{\widehat{Nq}_k}{NCW_k}\,(\%) \tag{18}$$

5. 실험 및 결과

5.1 파레토 분석 결과

Table 3에 단어 축소 과정에서의 계산결과들을 나타냈다. 설계조건(Design conditions)에서, $\overline{\alpha} = 1.161$는 설계조건으로 설정한 파레토 지수를 나타내고, NCW_k는 k에 해당하는 CW의 수를 나타내고, Nq_k는 설계조건 $q = 0.2$에 대해서 결정된 단어의 수를 나타낸다. 설계결과(Design results)에서, p_k, q_k, α_k 등은 주어진 설계조건에 대해서 계산한 결과를 나타낸다. 모든 k에 대해서 $\alpha_k > \overline{\alpha}\,(\overline{\alpha} = 1.161)$가 되어 설계조건을 만족하지 못하고 있다. 그래서 $p = 0.8$을 주고 $\widehat{Nq}_k$을 계산한 결과(Estimation results), 모든 k에 대해서 $\hat{p}_k \ge 0.8$, $\hat{q}_k \le 0.2$로 계산되어 $\hat{\alpha}_k \le \overline{\alpha}$의 조건을 만족하였다.

이상의 결과를 예로 설명하면, $k = 1$ (충돌사고)의 경우 해당하는 단어의 수 $NCW_k = 34$이고, 설계조건을 만족하는 단어의 수 $Nq_k = 6$인데, 평가결과 설계조건을 만족하지 못하였다. 그래서 추정한 결과 $\widehat{Nq}_k = 12$가 최적 단어의 수로 나타났고, 그 결과 $\hat{\alpha}_k \le \overline{\alpha}$의 조건을 만족하였다. 그래서 본 연구에서는 $\widehat{Nq}_k$에 해당하는 단어를 파레토 법칙을 따르는 사고종류별 최적의 최소 단어의 수로 결정하였다.

Table 3 Calculation results to k for the word reduction procedures

k	Design conditions			Design results			Estimation results			
	$\overline{\alpha}$	NCW_k	Nq_k	p_k	q_k	α_k	$\widehat{Nq}_k$	$\hat{p}_k$	$\hat{q}_k$	$\hat{\alpha}_k$
1	1.161	34	6	0.613	0.387	2.067	12	0.802	0.198	1.158
2	1.161	36	7	0.597	0.403	2.312	13	0.816	0.184	1.137
3	1.161	42	8	0.607	0.393	2.149	18	0.812	0.187	1.141
4	1.161	34	6	0.528	0.472	6.588	14	0.803	0.197	1.156
5	1.161	40	8	0.606	0.394	2.170	17	0.817	0.183	1.136
6	1.161	49	9	0.606	0.394	2.167	20	0.808	0.192	1.149
7	1.161	46	9	0.567	0.433	3.12	19	0.801	0.199	1.159
8	1.161	37	7	0.6	0.4	2.259	14	0.8	0.2	1.161
9	1.161	11	2	0.5	0.5	65535	6	0.808	0.192	1.148
10	1.161	56	11	0.593	0.408	2.401	24	0.804	0.196	1.155

5.2 파레토 차트를 이용한 검증

Fig. 3은 Table 3의 계산결과에 대한 유용성을 확인하기 위한 파레토 차트(chart)를 나타낸다. 파레토 차트는 해당 객체의 누적분포(%)를 값이 큰 순서부터 작은 순서로 나타낸 도표이다. Fig. 3은 $k = 10$ (아홉 가지 사고종류 전체)에 대한 파레토 차트에 $Fp_k(n_k) = 80\%$에 대한 객체의 수(그림에 NC로 표시)를 나타낸 것이다. Table 3의 $\widehat{Nq}_k\,(k = 10)$와 Fig. 3의 NC가 24로 일치하여 추정한 $\widehat{Nq}_k$가 유용함을 알 수 있다.

5.3 축소한 원인분류단어

Table 3의 $\widehat{Nq}_k$에 해당하는 $CW_{k,n_k}(n_k = \widehat{Nq}_k)$에 대한 CW_m의 인덱스 m을 Appendix A의 $ICW_{k,n_k}(n_k = \widehat{Nq}_k)$에서 구한 결과를 Table 4에 나타냈다. Table 4에 나타낸 m에 해당하는 원인분류단어는 위의 Table 2에서 확인할 수 있고, 이 단어들이 체계적인 주제어 구축에 적용하기 위한 최소의 단어가 된다. 그리고 Table 5에 식(18)으로 계산한 $\gamma_i(\%)$을 나타냈다.

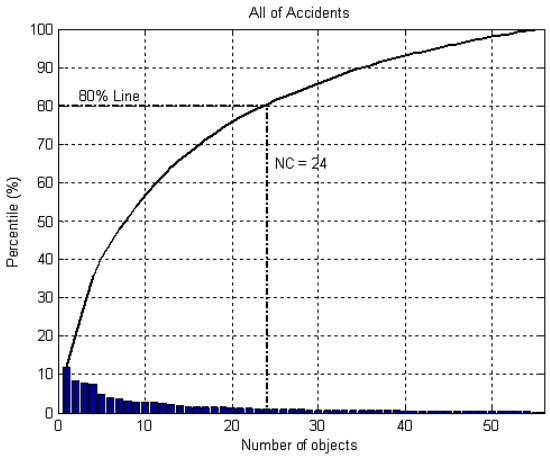

Fig. 3 Pareto chart to verify the number of estimated words $\widehat{Nq}_k\,(k=10)$ as shown in Table 3

Table 4 List of estimated the index m of $CW_{k,n_k}(n_k=\widehat{Nq}_k)$ corresponding to $\widehat{Nq}_k$ in Table 3

k	m	$\widehat{Nq}_k$
1	2, 3, 1, 6, 4, 10, 13, 9, 5, 12, 17, 35	12
2	1, 14, 3, 2, 4, 13, 5, 6, 12, 18, 10, 22, 21	13
3	1, 3, 4, 2, 9, 6, 17, 10, 12, 13, 5, 18, 19, 27, 7, 16, 21, 49	18
4	1, 3, 4, 18, 5, 13, 19, 29, 2, 12, 30, 6, 7, 49	14
5	1, 5, 4, 25, 32, 7, 2, 45, 43, 3, 6, 51, 8, 15, 17, 42, 53	17
6	4, 1, 2, 3, 24, 6, 7, 13, 5, 9, 8, 10, 15, 46, 11, 17, 23, 33, 12, 29	20
7	8, 7, 11, 1, 20, 10, 16, 3, 15, 37, 23, 28, 42, 38, 4, 34, 27, 29, 40	19
8	5, 1, 18, 3, 4, 9, 16, 48, 2, 6, 11, 12, 13, 15	14
9	5, 2, 1, 9, 31, 3	6
10	1, 2, 3, 4, 5, 6, 7, 8, 9, 13, 10, 11, 12, 18, 15, 17, 14, 16, 19, 29, 20, 22, 21, 30	24

Table 5 Calculation results for the reduction rates $\gamma_k(\%)$, where NCW_k and $\widehat{Nq}_k$ shows the number of target word and estimated word, respectively

k	NCW_k	$\widehat{Nq}_k$	$\gamma_k(\%)$
1	34	12	**64.7**
2	36	13	63.9
3	42	18	57.1
4	34	14	58.8
5	40	17	57.5
6	49	20	59.2
7	46	19	58.7
8	37	14	62.2
9	11	6	**45.5**
10	56	24	**57.1**
Average			**58.5**

Table 5에서 $k=1$ 이 $\gamma_k=64.7\%$ 로 최댓값을 나타내고, $k=9$ 에서 $\gamma_k=45.5\%$ 로 최솟값을 나타냈다. 특히, 사고종류 전체에 대한 $k=10$ 의 경우 $\gamma_k=57.1\%$ 로 나타났는데, 이 의미는 56개의 공통단어 중에서 파레토 법칙에 의거하여 추정한 24개의 단어만을 이용하더라도 공통단어의 80%를 설명할 수 있음을 나타낸다. 그리고 평균 축소율은 58.5%로 나타났다.

6. 결 론

본 연구에서는 주제어를 체계적으로 작성하는데 가장 중요한 사고종류별 공통된 단어를 최적의 최소 단어로 축소하여 도출하였다. 해양안전심판원의 정보포털에서 아홉 가지 사고종류에 대한 2,642개의 주제어를 획득한 후, 공통단어 56개를 도출해서 사고종류별 최적의 최소 단어를 추정하였다. 추정한 최소 단어는 파레토 차트(Pareto chart)를 이용하여 유효성을 입증하고, 축소율을 이용하여 축소 효율을 평가하였다. 연구결과를 요약하면 다음과 같다.

첫째, 추정한 최소 단어의 수와 파레토 차트로 분석한 최소 단어의 수가 일치하여 추정한 단어가 유효함을 확인하였다.

둘째, 축소율의 최대는 64.7%, 최소는 45.5%로 나타났고, 아홉 가지 사고종류 전체에 대해서는 57.1%로 나타났으며, 평균 축소율은 58.5%로 나타났다. 따라서 평균적으로 공통단어의 58.5%만 이용하더라도 사고의 80% 이상을 설명할 수 있는 최소 단어를 추정할 수 있었다. 이를 통해서 체계적인 주제어 구축 프레임 개발에 필요한 최소의 단어를 획득하였다.

셋째, 파레토 법칙을 이용하여 방대한 단어를 최적의 최소 단어로 축소할 수 있는 기법을 제안하였다. 제안한 기법은 단어로 구성된 데이터뿐만 아니라 다양하고 방대한 해양사고 데이터의 차원을 축소하는 경우에도 적용 가능할 것으로 고려된다.

추후, 본 연구에서 축소한 단어를 이용하여 주제어를 체계적으로 작성할 수 있는 프레임을 구축할 예정이다.

후 기

본 논문은 해양수산부의 '해양안전사고 예방시스템 기반연구(2단계)'과제의 연구결과임을 밝힌다.

References

[1] Armour Philip, Burkhauser R. V. and Larrimore Jeff(2014), "Using the Pareto distribution to improve estimates of topcoded earnings", white paper, NBER WORKING PAPER SERIES, Working Paper 19846, pp. 1-18, http://www.nber.org/papers/w19846.

[2] Brynjolfsson Erik, Hu Yu and Simester Duncan(2007),

"Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales", white paper, version November 2007, pp. 1-39, http://ebusiness.mit.edu/research/papers/2007.11_Brynjolfsson_Hu_Simester_Goodbye%20Pareto%20Principle_276.pdf.

[3] Cho S. S., Jang E. J. and Yim, J. B.(2015), "Research on the Numerical Data Construction for Marine Accidents," Proc. of Spring Seminar 2015, Korean Institute of Navigation and Port Research, pp. 193-195

[4] Fialova Alena, Jureckova Jana and Picek Jan(2004), "Estimating Pareto tail index based on sample means", REVSTAT - Statistical Journal, Vol. 2, No. 1, pp. 75-100.

[5] Finkelstein M., Tucker G. H. and Veeh J. A.(2006), "Pareto tail index estimation revised," North American Actuarial Journal, Vol. 10, No. 1, pp. 1-10.

[6] IMO(1997), CODE FOR THE INVESTIGATION OF MARINE CASUALTIES AND INCIDENTS, Resolution A.849(20) adopted on 27 November 1997, Appendix : Guidelines to assist investigators in the implementation of the Code.

[7] Jang E. J., Kang Y. M. and Yim J. B.(2016), "On the Analysis of Key Word in Korea Maritime Safety Tribunal to Prevent Human Error in Maritime Accidents", KAOSTS Joint Seminar 2016, Program book of Journal of Korean Navigation and Port Research, pp. 196-198.

[8] KMST(2003), "Machinery damaged accident for fishing ship No. 77 Dong-Myeong HO", Accident Analysis Report by Eastern KMST, No. 2003-001

[9] KMST(2007), "Final report 2007 for the analysis of judged prejudication of Korea Maritime Safety Tribunal", pp. 1-366.

[10] KMST(2014), 2014 Statistical Annual Report to Maritime Causalities (2008~2014 combined), Korea Maritime Safety Tribunal, pp. 1-118.

[11] KMST(2015), Web site for the Investigation and Judgement Information Portal of Maritime Causalities, http://data.kmst.go.kr/kmst/verdict/verdictAbstract/selectVerdictAbstract.do.

[12] KMST(2016), "Final report 2016 for the analysis of judged prejudication of Korea Maritime Safety Tribunal", Pub. reg. number 11-1192251-000012-01, pp. 1-57.

[13] MOF(2013), Law for the Investigation and Judgement of Maritime Causalities, No. 11690

[14] Rytgaard M.(1990), "Estimation in the Pareto distribution", Astin Bulletin, Vol. 20, No. 2, pp. 201-216.

[15] Sousa D. B. and Michailidis George(2004), "A Diagnostic Plot for Estimating the Tail Index of a Distribution", Journal of Computational and Graphical Statistics, Vol. 13, No. 4, pp. 1-22.

[16] Vilar-Zanon L. Jose and Lozano-Colomer Cristina(2007), "On Pareto conjugate priors and their application to large claims reinsurance premium calculation," Astin Bulletin, Vol. 37, No. 2, pp. 405-428.

[17] Wikipedia(2016), Pareto index, http://en.wikipedia.org/wiki/Pareto_index.

[18] Yim, J. B.(2009a), Development of Quantitative Risk Assessment Methodology for the Maritime Transportation Accident of Merchant Ship, Journal of Korean Navigation and Port Research, Vol. 33, No. 1, pp. 9-19

[19] Yim, J. B.(2009b), Implementation Techniques for the Seafarer's Human Error Assessment Model in a Merchant Ship: Practical Application to a Ship Management Company, Journal of Korean Navigation and Port Research, Vol. 33, No. 3, pp. 181-191

[20] Yim J. B., Yang W. J. and Kim H. T.(2014), Maritime Accident Analysis - Maritime Accident Analysis and Prevention in the View Point of Ship Operation, Jeilkihok, Mokpo, pp. 1-391.

Received 02 May 2017
Revised 21 June 2017
Accepted 23 June 2017

Appendix A Identified causation classification words in common words CW_m. In this table, ICW and CF denotes the index m in CW_m and cumulative frequencies according to m, respectively

n_k	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$		$k=6$		$k=7$		$k=8$		$k=9$		$k=10$	
	ICW	*CF*	*ICW*	*CF*	*ICW*	*CF*	*ICW*	*CF*	*ICW*	*CF*	*ICW*	*CF*	*ICW*	*CF*	*ICW*	*CF*	*ICW*	*CF*	*ICW*	*CF*
1	2	100	1	31	1	59	1	28	1	39	4	59	8	64	5	33	5	8	1	315
2	3	55	14	23	3	42	3	23	5	29	1	45	7	47	1	28	2	5	2	218
3	1	50	3	20	4	35	4	23	4	25	2	37	11	41	18	11	1	3	3	205
4	6	36	2	13	2	32	18	11	25	20	3	31	1	32	3	10	9	2	4	198
5	4	25	4	12	9	18	5	9	32	20	24	22	20	28	4	8	31	2	5	128
6	10	22	13	11	6	17	13	8	7	17	6	16	10	21	9	6	3	1	6	101
7	13	21	5	10	17	13	19	8	2	13	7	15	16	20	16	6	6	1	7	97
8	9	19	6	9	10	11	29	8	45	12	13	15	3	16	48	6	18	1	8	82
9	5	15	12	9	12	11	2	7	43	10	5	12	15	16	2	5	29	1	9	76
10	12	13	18	9	13	11	12	7	3	7	9	12	37	15	6	5	33	1	13	74
11	17	11	10	6	5	8	30	7	6	7	8	8	23	13	11	5	49	1	10	71
12	35	10	22	6	18	8	6	6	51	7	10	8	28	13	12	5			11	63
13	22	9	21	5	19	8	7	5	8	6	15	8	42	13	13	4			12	59
14	18	8	9	4	27	7	49	5	15	6	46	8	38	12	15	4			18	54
15	19	8	8	2	7	6	9	4	17	6	11	7	4	11	49	4			15	46
16	7	6	15	2	16	6	46	4	42	6	17	7	34	11	27	3			17	46
17	21	6	16	2	21	6	15	3	53	6	23	7	27	10	19	2			14	45
18	31	6	17	2	49	6	16	3	11	5	33	7	29	10	21	2			16	45
19	26	5	19	2	22	5	22	3	29	5	12	6	40	10	22	2			19	42
20	30	5	26	2	26	5	27	3	30	5	29	6	9	9	30	2			29	36
21	52	5	27	2	31	5	33	3	14	4	14	5	39	9	41	2			20	32
22	15	4	33	2	47	5	10	2	56	4	16	5	12	7	46	2			22	32
23	33	4	36	2	14	4	17	2	19	3	19	5	50	7	10	1			21	30
24	44	4	41	2	35	4	8	1	28	3	22	5	2	6	17	1			30	28
25	14	3	44	2	36	4	11	1	31	3	36	5	19	6	20	1			23	25
26	36	3	7	1	11	3	14	1	9	2	41	5	14	5	26	1			24	25
27	47	3	11	1	15	3	21	1	18	2	21	4	26	5	28	1			27	25
28	49	3	23	1	20	3	23	1	21	2	42	4	55	5	29	1			26	24
29	16	2	30	1	29	3	26	1	26	2	26	3	5	4	31	1			25	23
30	24	2	31	1	30	3	31	1	34	2	28	3	6	4	35	1			42	23
31	29	2	35	1	41	3	36	1	39	2	30	3	17	4	36	1			49	23
32	48	2	38	1	52	3	41	1	12	1	39	3	21	4	38	1			28	22
33	54	2	47	1	54	3	44	1	13	1	47	3	13	3	44	1			31	22
34	41	1	49	1	23	2	52	1	16	1	18	2	25	3	50	1			32	22
35			52	1	28	2			22	1	31	2	49	3	52	1			33	20
36			54	1	33	2			23	1	34	2	18	2	55	1			34	17
37					34	2			33	1	40	2	30	2	56	1			35	17
38					44	2			40	1	43	2	44	2					36	17
39					8	1			48	1	44	2	51	2					46	16
40					24	1			50	1	48	2	53	2					37	15
41					46	1					50	2	22	1					38	14
42					48	1					51	2	31	1					39	14
43											54	2	32	1					41	14
44											56	2	36	1					44	14
45											32	1	43	1					40	13
46											35	1	46	1					43	13
47											52	1							45	12
48											53	1							47	12
49											55	1							48	12
50																			52	12
51																			50	11
52																			51	11
53																			53	9
54																			54	8
55																			55	7
56																			56	7