

Orthographic Co-Reference Resolution Between Proper Nouns Through the Calculation of the Relation of "Replicancia"

Pedro AMO, Francisco L. FERRERAS, Fernando CRUZ and Saturnino MALDONADO

Dpto. de Teoría de la Señal y Comunicaciones

Universidad de Alcalá

Crta. Madrid-Barcelona km. 33,600

Alcalá de Henares-Madrid, SPAIN

{pedro.amo, tsflf, fernando.cruz}@alcala.es; smbascos@euitt.upm.es

Abstract

Nowadays there is a growing research activity centred in automated processing of texts on electronic form. One of the most common problems in this field is co-reference resolution, i.e. the way of knowing when, in one or more texts, different descriptions refer to the same entity.

A full co-reference resolution is difficult to achieve and too computationally demanding; moreover, in many applications it is enough to group the majority of co-referential expressions (expressions with the same referent) by means of a partial analysis. Our research is focused in co-reference resolution restricted to proper names, and we part from the definition of a new relation called "replicancia".

Though replicancia relation does not coincide extensionally nor intensionally with identity, attending to the tests we have carried out, the error produced when we take the replicantes as co-referents is admissible in those applications which need systems where a very high speed of processing takes priority.

1 Introduction

The proliferation of texts on electronic form during the last two decades has livened up the interest in Information Retrieval and given rise to new disciplines such as Information Extraction or automatic summarization. These three disciplines have in common the operation of Natural Language Processing techniques (Jacobs and Rau, 1993), which thus can evolve synergically.

Identification and treatment of noun phrases is one of the fields of interest shared both by Information Retrieval and Information Extraction. Such interest must be understood within the trend to carry out only partial analysis of texts so as to process them in a reasonable time (Chinchor et al., 1993).

The present work proposes a new instrument designed for the treatment of proper nouns and other simple noun phrases in texts written in Spanish language. The tool can be used both in Information Retrieval and Information Extraction systems.

1.1 Information Retrieval.

Information Retrieval systems aim to discriminate between the documents which form the system's entry, according to the information need posed by the user. In the field of Information Retrieval we can find two different approaches: information filtering -or routing- and retrospective, also known as *ad hoc*. In the first modality, the information need remains fixed and the documents which form the entry are always new. The *ad hoc* modality retrieves relevant texts from a relative static set of documents but, in contrast, admits changing information needs (Oard, 1996).

Information Retrieval systems make up representations of the texts and compare them with the representation of the information need raised by the user. The representation form most commonly used is a vector whose coordinates depend on the terms' frequency of occurrence, where such "terms" are elements of the text represented which can coincide with words, stems or words associations (n-grams) (Evans and Zhai, 1996).

At the present time designers try to enrich the sets of terms used in the representations with

noun phrases or parts of them as, for example, proper names. Thompson and Dozier conclude in (Thompson and Dozier, 1997) that: first, the recognition of proper nouns in the texts can be done with remarkable effectiveness; second, proper nouns occur frequently enough as well as in queries¹ to warrant their separate treatment in document retrieval, at least when working with case law documents or press news; third, the inclusion of proper nouns as terms in the documents' representations can improve information retrieval when proper nouns appear in the query.

As Gerald Salton has noted (Salton and Allan, 1995), in the field of information retrieval to determine the meaning of words is not as important as to ascertain if the meaning of the terms within the detection need coincide or not with the meaning of the terms in each document. When the terms are proper nouns we shall not talk about meaning but of reference, so that it will be interesting to know if the referents of the proper nouns included in the query coincide or not with the referents of the proper nouns which appear in each document.

Referents' resolution can not be done without an additional context (Thompson and Dozier, 1997). Nonetheless, in an Information Retrieval process we can not count on the additional information of a data base register, such as personal identification number, profession or age. Nor can we make a linguistic analysis as thorough as can be done in a language understanding system or in an Information Extraction system. Because of these reasons, the identification of proper nouns as co-references is usually confined to verify if the superficial forms of the two nouns under examination are close enough as to suppose that both of them refer to the same individual or entity.

1.2 Aim of the Work

Co-reference resolution is difficult because it demands the establishment of relationships between linguistic expressions -for example, proper names- and the entities denoted. So as to establish the co-reference between two expressions, we must identify first the referent of each one of them and then check if both coincide.

¹ "Query" is the translation of the user's information need into a format appropriate for the system.

In this work we propose a tool to resolve proper nouns co-reference, based only on the exam of those nouns' superficial form. To carry out this exam we have defined, in Section Two, the relation of *replicancia*. This relation links pairs of nouns -*replicantes*- which show certain resemblance between their superficial forms and use to have the same referent. In Section Three we propose an algorithm for the calculation of *replicancia*. This algorithm has the ability to learn pairs of *replicantes* and also allows to manually introduce pairs of proper nouns with the same referent and which register a high frequency of occurrence, although their orthographic forms do not bear any similarity at all. Section Four contains the results of the evaluation of co-reference resolution between proper nouns using only the relation of *replicancia*. Finally, in Section Five we draw some conclusions.

2.1 Definition of Replicancia

2.2 Antecedent

The need to develop an algorithm capable to resolve in a simple way to resolve proper nouns co-reference has been pointed-out by several authors. Among them, we can quote:

Dragomir Radev, in (Radev and McKeown, 1997), includes, between the forthcoming extensions of PROFILE system, an algorithm "...that will match different instances of the same entity appearing in different syntactic forms -e.g. to establish that 'PLO' is an alias for de 'Palestine Liberation Organization'..."

The second antecedent is (Bikel et al., 1998), where Bikel and his collaborators say about the future improvements of their own system:

"... We would like to incorporate the following into the current model: ...an aliasing algorithm, which dynamically updates the model (where e.g. IBM is an alias of International Business Machines)..."

2.2 Definition

We call *replicancia* to the relation which links proper nouns which presumably refer to the same entity if we attend exclusively to the nouns themselves, without paying attention to their respective contexts nor to which their actual referents may be. We shall call *replicantes* the nouns which maintain a relation of *replicancia* between them. We also assign the label

replicante² to the predicate which resolves the *replicancia* relation.

Replicancia is a diadic, reflexive, simetric, but not necessarily transitive relation. It is reflexive because every noun is a *replicante* of itself; simetric because if noun A is a *replicante* of noun B, noun B will also be a *replicante* of noun A; it is not transitive because the *replicancia* relation is established attending only to the nouns form, not to their referents, and it is possible that the same noun denotes two or more entities depending on its context of utilization. For example, let us say that noun A designates to José Salazar Fernández, noun B to JSF and name C to Juan Sánchez Figueroa; in this case A is a *replicante* of B and B a *replicante* of C, but A is not a *replicante* of C. Not being a transitive relation, the *replicancia* is neither an equivalence relation and does not induce a set of equivalence classes as it had been desirable (Hirsman, 1997).

We do not ask the predicate *replicante* to recognize as *replicantes* the pseudonyms, diminutives, nicknames, abbreviations and familiar names, although the most frequent tokens can be manually introduced as if they had previously been learned by the system.

Two proper nouns are *replicantes* when any of the following assumptions is satisfied:

- a) Both nouns or their canonic forms are identical³, as in *José María Aznar* and *Jose Maria Aznar*.
- b) One of them coincides with the initials of the other, be it in singular as in *Unión Europea* and *UE*, or in plural as *Estados Unidos* and *EE UU*. We also admit nouns with nexus as in the case of *Boletín Oficial del Estado* and *BOE*.
- c) The shorter noun is contained in the longer one and among the lexemes not shared there are no nexes. Under this rule it is admitted that *Julio Pérez de Lucas* is a *replicante* of *Pérez de Lucas*; nevertheless, *Caja de Madrid* is not a *replicante* of *Madrid*

because the part not shared, *Caja de*, includes the nexus *de*.

- d) Every word of the shorter noun, *N2*, is a version of some word included in the longer noun, *N1*, in the same relative order, although there can be words in *N1* which have no version in *N2*. A word *P2* is a version of another word *P1* when their canonic forms are identical, when *P2* is the initial of *P1* or when the initials of both coincide⁴. According to this fourth rule, the noun *JM Pacheco* is a *replicante* of *Juan Manuel Pacheco Fernández*; to verify it we compare the lexemes of the longer noun, one by one, with the lexemes which form the second name. In the comparison of *Juan* with *JM Pacheco* we identify the letter *J* as a version of *Juan*, after which remains *M Pacheco* as a rest. In the comparison between *Manuel* and *M Pacheco* we identify the letter *M* as a version of *Manuel*, after which remains *Pacheco* as the new rest that can be identified with the same lexeme in the first noun. As the new rest is empty, we decide that both names are *replicantes* of eachother although in the first noun there are lexemes left unmatched.

This definition of *replicante* makes it possible to identify the following list of nouns as *replicantes* of the first:

[*José Luis Martínez López*, *JL Martínez*, *J.L. Martínez*, *J Martínez*, *Luis Martínez*, *José Martínez*, *Martínez*, *JL M L*]

3 Replicancia Calculation Algorithm

Once the *replicancia* relation is defined, we are in a position to present the algorithm devised to calculate it. Although the algorithm has been developed in *Prolog*, we are going to expose it in a pseudocode, similar to that used in *PASCAL*, which uses the following notation: the symbols ":", "!" and "-" represent the conjunction, disjunction and negation of terms, respectively; the brackets "(...)" delimitate optional units; braces "{...}" are used to establish the prelation of logical operators; "::=" is the symbol chosen for

² We will use Courier New font to write the names of the predicates defined in Prolog which take part in the replicancia calculation algorithym.

³ The canonic form of a proper noun is its representation in small letters without accentuation.

⁴ In the last case the whole word *P1* in the longer noun is assimilated to only the initial of *P2*. The rest of the letters in *P2* must have some correspondence in the first noun so that we can consider both nouns as *replicantes*.

the definition and "==" refers to the assignation of values.

The calculation of the *replicancia* relation can impose a heavy computation burden if it is not somehow limited. The variety of forms in which two nouns can be mutually *replicantes* forces us to cover a long distance before being able to decide that two candidates are linked by such relation. That is why we have provided our algorithm with two instruments to reduce the computation burden; the first is its ability to learn, which permits the automatic creation of a data base of pairs of proper nouns mutually *replicantes*; the second is a filter which, based in a fast analysis of the initials of the nouns under comparison, reject most of the pairs of nouns that are not mutually *replicantes*. The main predicate can be expressed as follows:

```
replicante(N1,N2) ::=
  N1c = N2c /
  replicante_guardado(N1,N2) /
  filtro(N1,N2),
  resto_replicante(N1,N2),
  guarda_replicante(N1,N2).
```

where N_{xc} represents the canonic form of noun N_x and predicate *resto_replicante* is defined by:

```
resto_replicante(N1,N2) ::=
  siglas(N1,N2) /
  sin_prep(D12), {N2 = N1 - D12 /
  version(N1,N2)} /
  supprime_nexos(N1,N1s), {N2 =
  N1s - D12 / version(N1s,N2)}.
```

where N_x-N_y represents the lexemes sequency of noun N_x not present in noun N_y and $D12$ is $N1-N2$; predicate *sin_prep*(N_x) is true for the nouns N_x which do not include prepositions; *supprime_nexos* is the result of eliminating the noun of the lexemes in $N1$ which do not begin with a capital letter; finally, predicate *version*(N_x,N_y) is satisfied if every lexeme of N_y is a *version_palabra* of any of the lexemes of N_x without any alteration in the relative order of occurrence of each homologous lexeme, as we have already indicated in section 2.

In order to evaluate this algorithm we have designed the experiment described in Section 4.

4 Experiment and Results

We have defined the *replicancia* relation as an instrument for the resolution of co-reference between proper nouns, although we are perfectly aware of its limitations to identify as co-referentials nouns which are not linked by an orthographic relation, as occurs with nicknames and familiar names (*José* and *Pepe*, for example). In order not to conceal this limitations we have designed an experiment to evaluate the co-reference between proper nouns instead of evaluating the algorithm for the *replicancia* resolution, because this relation is only useful inasmuch it is capable to resolve co-reference between proper nouns.

In the context of natural language processing systems evaluation is very important to decide which collection of texts is going to be used. The current trend is not to use texts specifically prepared for the evaluation, but normal texts, that is to say, similar to the texts which the system will use in its normal operation. We have chosen documents available in the *World Wide Web*.

In our evaluation we will use a corpus manually processed composed by 100 documents with an extension scarcely over 1 MB, *HTML* code included. The documents come from five different newspapers, all of them spanish and available on electronic form. The newspapers, in alphabetical order, are: *ABCe*, *El Mundo*, *El País Digital*, *El Periódico On Line* and *La Vanguardia Electrónica*. The extension of the documents varies between 4kB and 25 kB, being the average around 10kB. The utilisation of these variety of sources to obtain the documents which form the corpus is very advisable, because people who work for the same newspaper tend to write similarly; the choosing of different document sources brings us closer to the style diversity characteristic of the texts available in the *World Wide Web*.

The result of co-reference resolution is the grouping of the document's selected objects -the proper nouns- in classes. The human analyst designs a *template* which includes the different instances of the same entity present in the document, and then compares it with the system's response, being this another grouping of objects in classes of co-reference. From this comparison we draw the system's quality evaluation.

The evaluation proceeding chosen is based on the measures used in *MUC* conferences (*MUC-7*,

1997) (Chinchor, 1997) (Hirschman, 1997). In MUC evaluations, the results of co-reference analysis are classified according to three categories:

COR - Correct. A result is correct when there is full coincidence between the contents of the template and the system's response.

MIS - Missing. We consider a result absent when it appears in the template but not in the answer.

SPU - Spurious. We consider a result spurious when it does not appear in the template although it is found in the system's response.

With the cardinals of these three classes we obtain two intermediate magnitudes which will be useful to calculate merit figures. These magnitudes are⁵:

POS - Possible. The number of elements contained in the template that contribute to the final scoring. It is calculated through the following expression:

$$POS = COR + MIS$$

ACT - Actual. The number of elements included in the system's response. It is calculated as follows:

$$ACT = COR + SPU$$

Once we have gathered the data relative to the classes of responses, we are prepared to calculate the measures of the system's quality. The metrics chosen are the ones normally used in Information Retrieval systems:

REC - Recall. A quantity measure of the elements of the answer-key included in the response of the system. Its expression is:

$$REC = \frac{COR + 0,5 \cdot PAR}{POS}$$

⁵ When the names of these measures appear in arithmetical expressions, they must be taken as the cardinals of the respective classes.

PRE - Precision. A measure of the quantity of elements of the response included also in the template. Its expression is:

$$PRE = \frac{COR + 0,5 \cdot PAR}{ACT}$$

F Measure - A compound measure introduced by van Rijsbergen (Rijsbergen, 1980). We use it with the control parameter $B=1$ to guarantee the same weight for recall and precision. Its general expression is:

$$F = \frac{(\beta^2 + 1) \cdot PRE \cdot REC}{\beta^2 \cdot PRE + REC}$$

From the cardinals of classes *COR*, *MIS* and *SPU* we obtain measures *REC*, *PRE* y *F*, the last one with parameter $B=1$. With all these data we fill in table I-1 of Annex I, where each line shows the metrics and data of a document. Table 4.1 shows overall results.

We have carried out the evaluation of the calculation of replicantes bearing in mind the use we intend to make of that relation. We have counted the cardinals of the co-reference classes among the proper nouns included in the text only in those classes whose elements adopted more than one form. Let us think, for example, in a text which had three classes of co-reference formed by the names⁶:

[(Madrid, 2)]
 [(José María Aznar, 1), (Aznar, 3)]
 [(Sumo Pontífice, 2)], (Juan Pablo II, 2),
 (Karol Wojtyła, 1)]

The first class refers to the entity in a single way; the second class includes two ways of referring to the entity and the third one three. For the evaluation of co-reference between nouns we consider that there are four references to the second entity and five to the third, which makes a total of nine nouns, although we know that the system is not prepared to identify as co-referentials the elements included in the third class. So, the result of the evaluation would be:

⁶ Each name shows the number of times it appears in the relevant text.

POS =9, ACT=9, COR=6, MIS=3, SPU=3

because of the nine elements subject of consideration, (POS), the system has recognized

Table 4.1. Summary of proper nouns coreference analysis

DOCUMENT	POS	ACT	COR	MIS	SPU	REC	PRE	F
AB1501	9	8	7	2	1	77,8	87,5	82,4
...
ZEDILLO	20	20	18	2	2	90,0	90,0	90,0
TOTAL	1523	1512	1296	227	216	85,1	85,7	85,4
Mean	15	15	13	2	2	84,6	86,2	85,0
Median	12	12,5	10	1	1	93,3	94	93,33
Std. Deviation	12,93	13	12	3,3	3,3	20,2	18,6	19,19
ABC	362	352	288	74	64	79,6	81,8	80,7
El Mundo	181	181	158	23	23	87,3	87,3	87,3
El País	434	432	377	57	55	86,9	87,3	87,1
El Periódico	350	354	289	61	65	82,6	81,6	82,1
La Vanguardia	196	193	184	12	9	93,9	95,3	94,6

nine (ACT) -six of them correctly located in their respective classes (COR), three located in two spurious classes (SPU) and three missing from their correct classes (MIS).

Briefly, we evaluate the resolution of co-reference between proper names without taking into account how the problem is solved by our algorithm; this way, the evaluation does not reflect the quality of the algorithm in calculating the replicancia, but the resolution of the coreference between proper nouns instead, which is the aim pursued.

5 Conclusion

The algorithm conceived does not calculate the co-reference, but the *replicancia* between proper names instead. *Replicancia* and co-reference do not coincide neither extensionally nor intensionally. As a consequence, referents of nouns linked by the replicancia relation are not bound by an identity relation, among other things because *replicancia* is not an equivalence relation, meanwhile coreference it is. The result is that one class of replicancia may contain names which refer to different entities and, however, not contain names which refer to the entity associated to the class.

Despite what it has been said, as the calculation of *replicancia* is much more simple than the calculation of coreferences, it is interesting, and even convenient, to calculate *replicancia* between nouns in limited contexts.

Table I-1 in Annex I shows the results of the experiment. Under the line showing the total

figures, we have added three lines with the average, median and mean deviation of the data obtained with each of the 100 documents analyzed. We can notice that the median take values between eight and nine points above average, which means that in most documents the co-reference between proper nouns is successfully decided. The negative burden comes from the variance, close to 25% of the recall and precision total values, which, along with the difference between average and median, forces us to think that most of mistakes are concentrated in some documents, in which precision and recall can be much smaller than expected.

In Figure 5.1 we have represented the histogram of the group of values obtained for F measure, values included in the right column in table I-1. It can be noted that more than half of the documents subject to evaluation obtain an F-measure over 0.9, and more than two thirds are over 0.8.

The statistical analysis of the system's quality measures is not enough to guarantee the representativity of the sample. It is also necessary to analyze its composition. The 100 documents of the sample were obtained aleatorily (random choice) during January 1998. The five sources used are different but all of them are relevant newspapers. Moreover, we have selected news under two different sections: domestic and international. The diversity of the documents which form the sample and the homogeneity of the measures obtained strengthen the hypothesis which states the results validity.

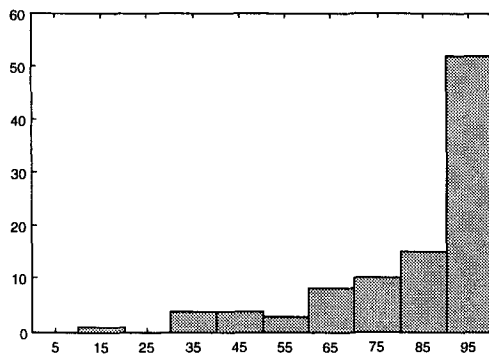


Figure 5.1 Histogram of the values of F-measure

The last five lines of table I-1, Annex I, register the total results distributed according to the documents origin.

The mistakes detected are mainly due to two causes: the presence of co-referents which are not replicantes and the presence of replicantes which are not co-referents, as sometimes it occurs with initials. Other errors, not attributable to the algorithm, are due to faults in the automatic system we have used for the extraction of proper nouns.

References

- Bikel, D., Miller, S., Schwartz, R. and Weischedel, R. (1998): "Nymble: a High-Performance Learning Name-finder". March 1998. In <http://xxx.lanl.gov/ps/cmp-lg/9803003>
- Chinchor, N. (1997): *MUC-7 Named Entity Task Definition. Dry Run Version. Versión 3.5.* Sep. 1997. In <http://www.muc.saic.com/>
- Chinchor, N., Hirschman, L. and Lewis, D. (1993): "Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)". *Computational Linguistics*, 1993, Vol. 19, N^o. 3.
- Evans, D. and Zhai, C. (1996): "Noun-Phrase Analysis in Unrestricted Text for Information Retrieval". May 1996. In <http://xxx.lanl.gov/ps/cmp-lg/9605019>
- Hirschman, L. (1997): *MUC-7 Coreference Task Definition. Versión 3.0.* Julio 1997. In <http://www.muc.saic.com/>

- Jacobs, P. and Rau, L. (1990): "SCISOR: Extracting Information from On-line News". *Communications of the ACM*, noviembre 1990, Vol. 33, N^o.11.
- MUC-7 CFP(1997): "Seventh Message Understanding System and Message Understanding Conference (MUC-7). Call For Participation". 1997. In <ftp://ftp.muc.saic.com/pub/MUC/participation/call-for-participation>
- Oard, D. W. (1996): *Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications.* PhD thesis, University of Maryland, College Park, August 1996. <http://www.ee.umd.edu/medlab/filter/papers/thesis.ps.gz>.
- Radev, D. and McKeown, K. (1997): "Building a Generation Knowledge Source using Internet-Accessible Newswire". Feb. 1997. En <http://xxx.lanl.gov/ps/cmp-lg/9702014>
- Rijsbergen, C. (1980): *Information Retrieval.* Butterworths. Londres, 1980.
- Salton, G. and Allan, J. (1995): "Selective Text Utilization and Text Traversal". *Int. J. Human-Computer Studies* (1995) 43, pp 483-497.
- Thompson, P. and Dozier, C. (1997): "Name Searching and Information Retrieval". June 1997. In <http://xxx.lanl.gov/ps/cmp-lg/9706017>

Appendix I. Evaluation data

Table I- 1 Results of co-reference between proper names resolution⁷

D	POS	ACT	COR	MIS	SPU	REC	PRE	F	D	POS	ACT	COR	MIS	SPU	REC	PRE	F
1	9	8	7	2	1	77,8	87,5	82,4	51	12	12	9	3	3	75,0	75,0	75,0
2	31	29	24	7	5	77,4	82,8	80,0	52	10	10	9	1	1	90,0	90,0	90,0
3	9	9	8	1	1	88,9	88,9	88,9	53	6	6	5	1	1	83,3	83,3	83,3
4	0	0	---	---	---	---	---	---	54	23	23	23	0	0	100	100	100
5	0	0	---	---	---	---	---	---	55	10	9	9	1	0	90,0	100	94,7
6	4	1	1	3	0	25,0	100	40,0	56	20	20	8	12	12	40,0	40,0	40,0
7	0	0	---	---	---	---	---	---	57	19	17	12	7	5	63,2	70,6	66,7
8	11	10	7	4	3	63,6	70,0	66,7	58	27	27	24	3	3	88,9	88,9	88,9
9	2	2	2	0	0	100	100	100	59	6	6	6	0	0	100	100	100
10	11	11	10	1	1	90,9	90,9	90,9	60	12	12	7	5	5	58,3	58,3	58,3
11	16	15	2	14	13	12,5	13,3	12,9	61	15	16	6	9	10	40,0	37,5	38,7
12	15	15	15	0	0	100	100	100	62	2	2	2	0	0	100	100	100
13	12	12	10	2	2	83,3	83,3	83,3	63	6	6	6	0	0	100	100	100
14	7	5	4	3	1	57,1	80,0	66,7	64	29	29	20	9	9	69,0	69,0	69,0
15	4	4	4	0	0	100	100	100	65	32	32	32	0	0	100	100	100
16	5	5	5	0	0	100	100	100	66	2	3	2	0	1	100	66,7	80,0
17	15	15	15	0	0	100	100	100	67	5	5	5	0	0	100	100	100
18	25	25	12	13	13	48,0	48,0	48,0	68	89	89	85	4	4	95,5	95,5	95,5
19	5	4	2	3	2	40,0	50,0	44,4	69	14	15	6	8	9	42,9	40,0	41,4
20	18	18	18	0	0	100	100	100	70	23	23	21	2	2	91,3	91,3	91,3
21	21	21	21	0	0	100	100	100	71	35	35	33	2	2	94,3	94,3	94,3
22	5	5	5	0	0	100	100	100	72	7	7	7	0	0	100	100	100
23	9	9	9	0	0	100	100	100	73	7	8	7	0	1	100	87,5	93,3
24	18	20	17	1	3	94,4	85,0	89,5	74	21	20	10	11	10	47,6	50,0	48,8
25	24	24	20	4	4	83,3	83,3	83,3	75	36	37	26	10	11	72,2	70,3	71,2
26	12	13	10	2	3	83,3	76,9	80,0	76	30	30	28	2	2	93,3	93,3	93,3
27	17	17	17	0	0	100	100	100	77	50	50	47	3	3	94,0	94,0	94,0
28	19	19	18	1	1	94,7	94,7	94,7	78	30	28	28	2	0	93,3	100	96,6
29	4	4	3	1	1	75,0	75,0	75,0	79	26	26	25	1	1	96,2	96,2	96,2
30	5	5	3	2	2	60,0	60,0	60,0	80	12	11	11	1	0	91,7	100	95,7
31	15	15	15	0	0	100	100	100	81	2	2	2	0	0	100	100	100
32	15	15	15	0	0	100	100	100	82	12	12	12	0	0	100	100	100
33	22	24	15	7	9	68,2	62,5	65,2	83	3	3	1	2	2	33,3	33,3	33,3
34	15	14	13	2	1	86,7	92,9	89,7	84	20	20	15	5	5	75,0	75,0	75,0
35	44	45	35	9	10	79,5	77,8	78,7	85	3	3	2	1	1	66,7	66,7	66,7
36	31	32	26	5	6	83,9	81,3	82,5	86	14	14	14	0	0	100	100	100
37	48	48	47	1	1	97,9	97,9	97,9	87	13	13	13	0	0	100	100	100
38	8	8	8	0	0	100	100	100	88	4	2	2	2	0	50,0	100	66,7
39	29	27	19	10	8	65,5	70,4	67,9	89	8	8	8	0	0	100	100	100
40	15	15	15	0	0	100	100	100	90	9	9	9	0	0	100	100	100
41	16	18	16	0	2	100	88,9	94,1	91	6	6	6	0	0	100	100	100
42	19	19	15	4	4	78,9	78,9	78,9	92	4	4	4	0	0	100	100	100
43	19	19	16	3	3	84,2	84,2	84,2	93	2	2	2	0	0	100	100	100
44	18	18	16	2	2	88,9	88,9	88,9	94	17	17	17	0	0	100	100	100
45	10	10	6	4	4	60,0	60,0	60,0	95	10	10	10	0	0	100	100	100
46	11	8	8	3	0	72,7	100	84,2	96	8	8	8	0	0	100	100	100
47	5	5	4	1	1	80,0	80,0	80,0	97	16	16	16	0	0	100	100	100
48	20	20	20	0	0	100	100	100	98	2	2	2	0	0	100	100	100
49	7	7	6	1	1	85,7	85,7	85,7	99	5	5	5	0	0	100	100	100
50	29	30	27	2	3	93,1	90,0	91,5	100	20	20	18	2	2	90,0	90,0	90,0

⁷ Each line stands for a document