TDT4215
Web-intelligence project report

Arnfinn Gjørvad, Bjørn Henninen, Magnus Bae

2013-04-23

> *Smart behavior in the Web comes from smart applications on the Web, not from the infrastructure.*
>
> *- Dean Allemang, James Hendler. 2011*

Number of words: 2387

Contents

Introduction

In this project we've built a system that analyzes patient cases, and finds and presents relevant chapters from the Norwegian medical handbook. The system works by first creating an index of the handbook's various chapters, which is then improved with data from the Norwegian version of the ICD10 ontology, and ATC ontology. Afterwards the cases are parsed and queries are made to Apache Lucene based on the parsed cases to find relevant chapters. The user is then presented with the top results from the query in a textual manner.

In this report we will take a brief look on the system, its components, and the technology behind it.

System architecture

The system is built as a standalone java application. This makes it easy to develop and test the system, and it's also quite easy to port to EJB or JSF (server-side Java technologies). Java was chosen as the development language as it supported all the technologies the group wanted to utilize in the project and the group was more competent in Java then in other programming languages. Since Lucene is written in java, integration is also made significantly easier. Theoretically our choice of Java should speed up the development process.

Components

The system is built using traditional object models, with a GUI frontend built with SWT. There's not much functionality in the GUI so the application is very portable if one would want to create a web-app or something similar.

Package model

The system is divided into the following packages:

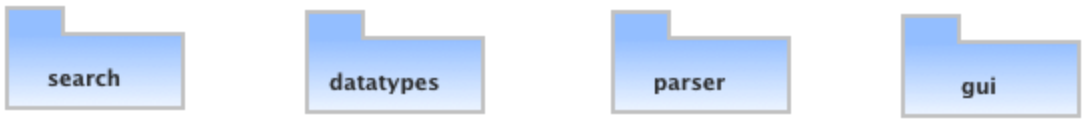

Figure 1 - Package diagram

Class model

The classes in the different packages interact with each other, this diagram shows all interaction, between packages and within packages without differentiating between the different packages.

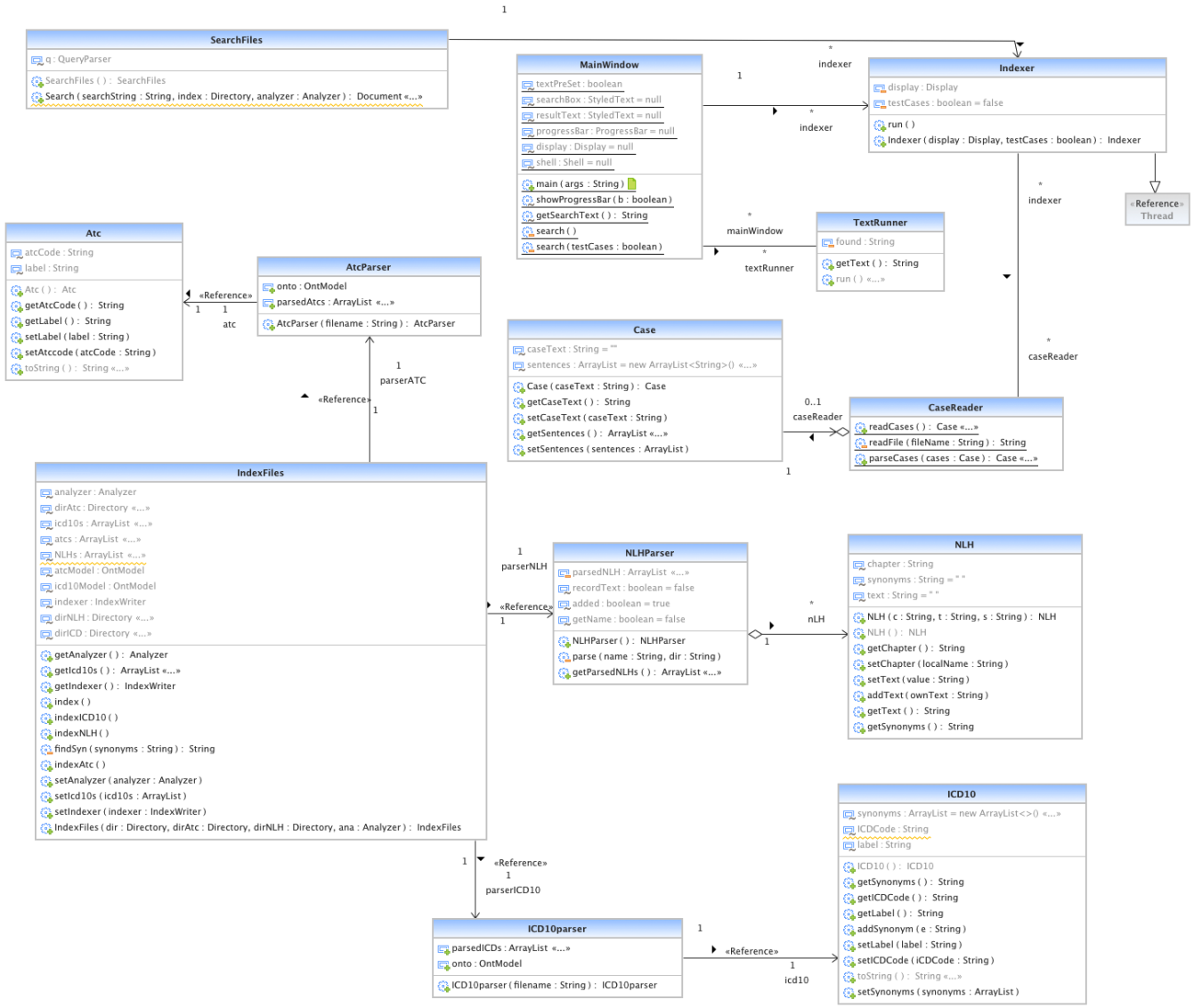

Figure 2 - Class model

Control flow

At the start of the application the GUI is initialized and waits for user interaction. On the first run no search index is built and when the user starts a search the GUI starts a new thread which first checks if the index exists, when it realizes it doesn't it starts by parsing the ATC ontology, and then the ICD10 before it finally parses the NLH. The order of the parsing was important, since the NHL used data from the two other sources in its own parsing. For each entity in the

NHL, we search the two other sources for relevant entries. If we find anything, we add the code, the name and the synonyms to the entities synonyms. Afterwards the actions are the same as if the index already existed and could be opened. The program queries the index and gives the user an output.

To keep the user informed on the progress there's a progress bar showing during search and indexing. This way the user knows that the application isn't frozen and is working. The progress during index-building is a bit of guess-work though as the time it takes cannot be estimated properly and we cannot monitor the progress of the process. The progress bar increases with 1% every 200ms until it reaches 94% where it stops and waits for the search to continue. On the computers we've tested on this seems to be fairly close the real progress.

Theories

Vector models
Vector models allow partial matches, which is a huge improvement over a regular Boolean search. Its particularly important for this assignment, since so much of the text in the patient cases are irrelevant and doesn't produce any good search results. The group chose to work with Apache Lucene as it supports vector models, is mature and should significantly shorten the development time compared to building something from scratch.

Stemming
Stemming is the manipulation of words so that they are close to or in their root form. To achieve stemming the group chose to use an analyzer created for the Norwegian language in conjunction with Lucene. The analyzer was provided through the Lucene project.

Results

In this project the group has built a system that analyzes patient cases and finds and presents relevant chapters from the Norwegian medical handbook. To aid in the work the groups were provided with a gold standard which can be used for evaluation. [1]

The system indexes the ATC and Norwegian ICD10 ontologies, and the Norwegian Medical Handbook (NLH). Using the ATC and ICD10 ontologies the system generates synonyms for different words and using these synonyms evaluates the search queries. We also save the name of the main chapter and add it is a synonym (e.g. for T2.2.1, we would add the name of T2.2, “Kreftsykdommer”).

The system can also calculate recall and precision automatically for the example cases. This is possible since we stored the chapter numbers from the gold standard.

Below are some tables showing the results in the various iterations of the project.

[1] Document from It’s Learning: Gold Standard

Case	Result	Recall (top 4)	Precision (top 4)
Case 1	T3.1 T3.1.3 L3.1 L3.2	50% 1/2	25% 1/4
Case 2	T10.3 T8.4 L24.2 L1.7.1	50% 1/2	25% 1/4
Case 3	T1.10 T1.14.2 T1.14 L1.2	100% 1/1	25% 1/4
Case 4	T8.3 T8.3.2 T8.4 T3.1	66% 2/3	50% 2/4
Case 5	T2.2 T1.4 T1.8 L1.2	0% 0/3	0% 0/4
Case 6	T11.3 T1.7 L1.2 L1.2.11	0% 0/4	0% 0/4
Case 7	T20.1 T20.1.2 L4.5 L3.2	0% 0/4	0% 0/4
Case 8	T11.3 T11.1 T10.2 L1.2	0% 0/4	0% 0/4
Overall		5/23 ~22%	5/32 ~15%

Table 1 - Recall and precision when not indexing synonyms.

Case	Result	Recall (top 4)	Precision (top 4)

Case 1	T3.1 T3.1.3 L3.1.1 L3.1	100% 2/2	50% 2/4
Case 2	T10.3 T10.2 L3.2.5 L1.7.1	50% 1/2	25% 1/4
Case 3	T1.10 T6.2 T1.14.2 L1.2	100% 1/1	25% 1/4
Case 4	T8.3 T8.3.2 T8.3.1 T2.2.3	100% 3/3	75% 3/4
Case 5	T2.2 T12.10.1 T4.1 L3.9.1	33% 1/3	25% 1/4
Case 6	T2.2 T11.3.2 L1.2.3 L1.2.1	25% 1/4	25% 1/4
Case 7	T20.2.3 T21.1.1 T20.2 L20.1	25% 1/4	25% 1/4
Case 8	T11.3.2 T15.3 T11.3 L2.1.4	25% 1/4	25% 1/4
Overall		11/23 ~48%	11/32 ~34%

Table 2 - Recall and precision when synonyms are being indexed and utilized in search, but without any weighting of the different fields.

Case	Result	Recall (top 4)	Precision (top 4)
Case 1	T3.1 T3.1.3	100% 2/2	50% 2/4

	L3.1.1 L3.1		
Case 2	T10.2 T10.3 T8.4 L1.7.1	50% 1/2	25% 1/4
Case 3	T1.10 T1.14.2 T1.14 L1.2	100% 1/1	25% 1/4
Case 4	T8.3 T8.3.2 T8.3.1 T3.1	100% 3/3	75% 3/4
Case 5	T2.2 T4.1 T12.10.1 T8.4	33% 1/3	25% 1/4
Case 6	T11.3.2 T11.3 T2.2 L1.2	50% 2/4	50% 2/4
Case 7	T20.2.3 T21.1.1 L20.1 L20.1.2	25% 1/4	25% 1/4
Case 8	T11.3.2 T11.3 T11.1 L1.2	25% 1/4	25% 1/4
Overall		12/23 ~52%	12/32 ~37%

Table 3 - Recall and precision when synonyms from ICD10 and ATC, as well as some data about the parents in the NHL, are being indexed and utilized in search, with optimized weighting determined using empirical methods.

Visibly, the biggest improvement was achieved when adding in data from the other two sources. This upped the recall rate from 22% to 48%. Over a 100% improvement. Adding in some data about sub-classes and weighting the fields used in the search only gave a meager 4 pp. improvement, to 52%. However, it did help with subjective accuracy, as the group felt the results obtained were closer to the gold standard.

To find the most optimal weighting we let the system iterate over different possible weights which we knew was within the range that would provide the best results. The weighting for the synonyms that gave the best recall and precision was 0.4.

With regards to performance the system is very fast and responsive once the index is built (less than one second response-time). Building the index usually takes about 10-20 seconds depending on the system.

Discussion

In this section we'll discuss limitations in project, system, and results.

Accuracy

Even though the mathematical comparison of our results with the gold standard seemingly shows low accuracy, the group would like to argue that our system is quite accurate and produces relevant results. In some cases the gold standard asks for a sub-sub-chapter like 3.2.2, while our system produces 3.2 instead. This disqualifies our system for being correct according to the gold standard, but it's very close to it, and the result is highly relevant as well. Also, since the system by default gives 4 results, in the cases where the gold standard only calls for one result this also leads to lower precision.

When calculating recall and precision we did not compare the position in which the gold standard wants the results to be in, only if they were in the search results or not. However, we can manually observe that in most cases our order is quite close to the gold standard. And the top result is relevant in 7 of the 8 cases.

GUI

The GUI created for this assignment is in no way meant to be an example of good coding practices. The GUI is simply there for convenience and proof-of-concept and little work went into creating it.

Portability

The system being created in Java is very portable across platforms. During development the system has been tested on both OS X and Windows. The only requirements for the system is SWT support and JRE 1.7 (Java SE 7) or above. The delivery comes with pre-built versions for

Windows with both 64-bit and 32-bit JVM and also for OS X with a 64-bit JVM. Tested on Windows 7, Windows 8, and OS X Mountain Lion (10.8). SWT builds can be downloaded for other systems as well, it should be enough to replace the swt.jar in the *nlhparser_lib* folder with the one for actual architecture. SWT can also be built from source.[2]

Issues

Working on this project has not been trouble-free. There have been problems with both the availability of resources and documentation pertaining the version of Lucene we chose to work with.

As for resources some of the resources that we were supposed to use became available online quite late and resulted in a bit of a time constraint on the project. We were also given an turtle ontology of the Norwegian extension of ATC, which we weren't able to parse programmatically in Jena, even though the file looked OK in Protege (after removing errors in the file by recipe from other students in the forum on It's learning). We do think that if we'd been able to parse this file we'd get better results than we were able to get with the international ATC.

As for Lucene it took some time to get started with Lucene 4 since all examples and most documentation was for version 3.x. Trying to follow these examples often resulted in outcomes that wasn't as expected.

Future work and improvements

If the system is to be developed further the most natural things to do is the following:

1. Parse and store chapter URL's in the index
    a. Make the results dynamic by including the URL's
2. Make the results more dynamic by making them clickable to get the chapter text (with some nice formatting)
3. If the system is to be provided as a web-service, create the index in a singleton-bean or similar singleton-pattern.
4. Results can easily be formatted as JSON, XML, or other type of data-format if wanted. Right now, however, the results are being propagated to the GUI as strings.

Some things could also be done to improve the results even further. While we use data about parents in the NHL we do not use information about the relationships in the ICD10 or ATC. The best results might have been obtained by using a semantic web, but even something as simple as adding the name of the parent node as a synonym (or as another weighted field) might have improved the results some.

Conclusions

[2] http://eclipse.org/swt/

In this report we have looked at the system built by our group in TDT4215. The system parses medical codes, words, and the Norwegian Medical Handbook (NLH) and allows users to find relevant chapters in the handbook by entering a written patient case. The patient cases are written and usually contain several lines of text, or even multiple paragraphs. To analyze these properly is quite a task for a computer system.

Looking at the results we see that our system, even though it does not quite meet the gold standard, makes a fairly good case. Our system provides consistently relevant search results, even though some irrelevant does sneak into the search results. We would like to emphasize that our system is by no means an expert-system and is only meant to be a helpful utility to automatically suggest relevant pages in the medical handbook, not to decide wheter or not the chapters actually are relevant.

Working with this project the group has learned a lot about working with ontologies, parsing, searching, and Apache Lucene. This is knowledge and experience we think and hope will be useful in the future, both for school and work. There has also been a fair share of trouble, most of which has been due to lack of documentation and examples for Lucene 4, but all in all we are quite happy with the result.

Appendix 1: Subtasks

Clinical note	Sentence	ATC
1	1 2 3 5	A10 G03BA G03BB V04CA A10AB02 A10X A10 V04CA A10AB02 DB00030 A10AF01 A10AB02 DB00030 A10AF01
2	4 14 15 17 22 24 25 27	A10BH ,G03BB ,V03AB27 ,G03BA , A10BH ,G03BB ,V03AB27 ,G03BA , J07BL01 ,V04CE ,J07AN01 ,A05BA , C10AX06 ,G03BA ,G03BB , J06BA01 ,J06BA02 ,J06BA , R07AA , DB00069 ,L03AB09 ,C04AA ,R07AA , R03AC ,D01AE06 ,G03BA ,G03BB ,
3	2 4 7 8 16 17 20 21	N04AB ,M03BC , A11GA01 ,C01AA06 ,B01AD12 ,DB00126 , J07BM01 ,J07BM02 , A11GA01 ,C01AA06 ,B01AD12 ,DB00126 , D01AE06 ,DB00417 ,C01AC01 ,DB01053 , DB00829 ,N05BA01 , A11GA01 ,C01AA06 ,B01AD12 ,DB00126 , N04AB ,M03BC ,
4	1 2 9	R03AC ,C04AA ,R03CC ,D01AE06 , C10AX06 ,G03BA ,G03BB , DB00727 ,
5	2 7 9 12	G04CB ,C10AX07 ,G03BB , J06 ,J06AA05 ,J06AA ,J06A , B05AA10 ,B05AA08 ,B05AA09 , J06BA01 ,J06BA02 ,J06BA ,
6	1 2 4 5 6	N05AA V04CJ ,V04CE ,V04CA ,V04B , J01CE ,J01C ,DB00417 ,DB01053 , J01CE ,J01C ,DB00417 ,DB01053 , A10AD ,A10AB ,
7	13 14	R03AC ,D01AE06 ,A11HA03 ,DB00163 , A11HA03 ,DB00163 ,

	15 17 18 21 24	G03BB ,N02AX ,N01AH ,N02A , D01AE06 ,DB00069 ,L03AB09 ,C04AA , R03AC ,C04AA ,R03CC ,D01AE06 , A10AD ,A10AB , A06AH ,N02AX ,N01AH ,N02A ,
8	5 8 9 11 16	C10AX06 G03BA G03BB J01CE J01C DB00417 DB01053 DB00069 G04CB C10AX07 G03BB A11GA01 C01AA06 B01AD12 DB00126 C01AA06 B01AD12 DB00126 C02N

Case number: 1

Sentence number: 1
P702, O240, E10-E14, H280
Sentence number: 2
E10-E14, Z713, R730, E232
Sentence number: 3
D485, T801, T802, O84
Sentence number: 4
P702, O240, O241, P051
Sentence number: 5
P702, R11, Z450, Z305
Sentence number: 7
O366, P035
Sentence number: 8
E501, E500, L853, E502
Sentence number: 10
Z348, Z340, Z34, Z013
Sentence number: 11
Z386, Y872, Z380, Z383

Case number: 2

Sentence number: 1
Q113, N501, Z701, M232
Sentence number: 2
Z566, Z565
Sentence number: 3
N258, T838, P124

Sentence number: 4
P911, N183, N184, P521
Sentence number: 5
Q383, C029, K148, K145
Sentence number: 6
D24, N641, Q831, O11
Sentence number: 7
R042, P241, P924, R05
Sentence number: 8
N63, D24, N641, Q831
Sentence number: 9
R462
Sentence number: 10
N63, D24, N641, Q831
Sentence number: 11
Z711, N951, N953, E848
Sentence number: 12
C040, M321, H702, H302
Sentence number: 13
Z598, Z600, R462, Z721
Sentence number: 14
N183, N184, P521, J441
Sentence number: 15
J459, J45, J450, J451
Sentence number: 16
C250, C251, C252, C080
Sentence number: 17
A502, A500, A662, N183
Sentence number: 18
A161, Z044, Z028, Z048
Sentence number: 19
R85, R86, R944, R191
Sentence number: 21
Y900, I441, I440, Q923
Sentence number: 22
P720, R01, R012, R011
Sentence number: 23
T318, T328, N920, N923
Sentence number: 24
A065, C780, Q333, D860
Sentence number: 25
Q333, D860, Y630, P051
Sentence number: 26
O346, O345, O344, O348
Sentence number: 27
Z348, Z340, D103, P051

Case number: 3

Sentence number: 1
H546, H549, G521, O324
Sentence number: 2
Q981, S982, D817, D816
Sentence number: 3
E668, Z763, Z721, I495
Sentence number: 4
I495, T07, R51, R074
Sentence number: 5
Z621, J110, J100, J09
Sentence number: 6
Z636, Z611, Z750, Z553
Sentence number: 7
C925
Sentence number: 8
P073, O601, E301, B171
Sentence number: 9
B354, Z028, O074, Z048
Sentence number: 11
D235, C792, R202, D485
Sentence number: 12
P000, T317, T327, Z013
Sentence number: 13
G02, G038, Z390, Z036
Sentence number: 14
G970, C110, O074, Y847
Sentence number: 15
Z765
Sentence number: 16
O603, O60, O600, U80
Sentence number: 17
C040, C041, H702, H302
Sentence number: 18
S840, D227, S819, Z991
Sentence number: 19
R85, A415, R98, R827
Sentence number: 20
A013, A160, B171, A032
Sentence number: 21
S982, T022, T023, Z373
Sentence number: 22
M022, Z637, T881, T880

Case number: 4

Sentence number: 1
Z574, J69, T512, O96

Sentence number: 2
N63, D24, N641, Q831
Sentence number: 3
C050, R26, R260, R261
Sentence number: 4
T310, T320, Z720, J684
Sentence number: 5
R633, P924, N184, J690
Sentence number: 6
H912, R96, R961, R960
Sentence number: 7
T315, T325, N881, M232
Sentence number: 8
I200, I208, I201, I209
Sentence number: 9
Z380, Z383, O324, N922

Case number: 5

Sentence number: 1
N914, M232, N911, N818
Sentence number: 2
N185, R462, R15, R195
Sentence number: 3
R790, Z762, R15, R195
Sentence number: 4
I84, O224, I840, Z751
Sentence number: 5
Z875, O074, Z048, Z100
Sentence number: 6
S610, N921, S600, N926
Sentence number: 7
I847, I843, I844, I845
Sentence number: 8
R15, C846, R762, R195
Sentence number: 9
D564, D573, Y847, P050
Sentence number: 10
S046, H933, C724
Sentence number: 11
M061, G523, Q564, Q56
Sentence number: 12
Q409, J069, H721, Y847
Sentence number: 13
C189, Z048, D126, Z701

Case number: 6

Sentence number: 1
Q845, J351, R59, R599
Sentence number: 2
C866, Z32, R762, C846
Sentence number: 3
I252, B909
Sentence number: 4
O601, O603, O60, O600
Sentence number: 5
Q104, Z911, Y69, Y633
Sentence number: 6
J698, Z574, Z575, J69
Sentence number: 7
G820, P293, R53, G810

Case number: 7

Sentence number: 1
S200, N64, A662, C509
Sentence number: 2
C798, C795, S722, M966
Sentence number: 3
Z701, K912, Q656, K913
Sentence number: 4
M796, R101, R522, R529
Sentence number: 5
R529, D168, S321, C414
Sentence number: 6
R788, R85, R98, R491
Sentence number: 7
R190, O331, C495, C763
Sentence number: 8
M70, R101, R522, R529
Sentence number: 9
O688, R101, R522, R529
Sentence number: 10
O31, Z022, G405, C542
Sentence number: 11
M146, R101, R522, R529
Sentence number: 12
Z550, Z356, N140, Z701
Sentence number: 13
Q992, Z270, Z272, Z273
Sentence number: 14
O311, E560, Q738, B962
Sentence number: 15
O140, P210, Z269, K904
Sentence number: 16

O810, P071, P070, R680
Sentence number: 17
T512, H544, H545, C931
Sentence number: 18
P521, Z036, T512, T887
Sentence number: 19
S819, O13, E163, R700
Sentence number: 20
Z550, O342, Z721, R11
Sentence number: 21
J698, Z574, Z575, J69
Sentence number: 22
G252, K121, H258, T748
Sentence number: 23
R11
Sentence number: 24
T758, T97, Q381, P026

Case number: 8

Sentence number: 1
S046, H933, C724, O324
Sentence number: 2
Z412, D816, E550, K732
Sentence number: 3
Z621, R509, A78, R508,
Sentence number: 4
Z631, Z620, Z621, Z711
Sentence number: 5
N183, P521, P911, Z490
Sentence number: 6
C098, C024, C099, D104
Sentence number: 7
C770, C771, C773, R599
Sentence number: 8
O601, O603, O60, O600
Sentence number: 9
S92, Z515, P210, H544
Sentence number: 10
Z710, T313, T323, S011
Sentence number: 11
Z028, A013, Z048, B171
Sentence number: 12
Z044, D806, Z028, Z048
Sentence number: 13
Q353, Q375, Q373, C051
Sentence number: 14
H530

Sentence number: 15
K122, Z701, Q301, J340
Sentence number: 16
D589, D593, B171, E782