Reconstructing Corrupt DEFLATEd Files

By

Ralf Brown

Presented At

The Digital Forensic Research Conference

DFRWS 2011 USA New Orleans, LA (Aug 1st - 3rd)

Reconstructing Corrupt DEFLATEd Files

Ralf D. Brown
Carnegie Mellon University

3 August 2011

Why do we care about DEFLATE compression?

DEFLATE is Ubiquitous

- Many file types are in fact ZIP archives:
  - OOXML (.docx, .xslx, .pptx)
  - OpenDocument (.odt, .odp, .odg, .ods)
  - ePub e-books, Comic Book archives (.epub, .cbz)
  - Java applications and Android apps (.jar, .apk)
  - WinAmp and Tribe 2 skins (.wsz, .vl2)
- Numerous other compressors use DEFLATE:
  - gzip
  - zlib
  - ALZip

Off-the-Shelf ZIP Recovery Programs

- Can list archive contents based on central directory and/or scanning for local file headers
- Can extract intact archive members
- May be able to extract truncated members
- Can NOT extract members whose beginning is missing or overwritten
- Can NOT deal with split archives where one or more segments are missing

Introducing ZipRec

- Prototype program to extract files from ZIP archives
  - Full recovery of intact members
  - Partial recovery of truncated members
  - Partial recovery from members missing beginning
  - Partial recovery from members with missing or corrupted middle
- Also offers some support for gzip files and zlib streams

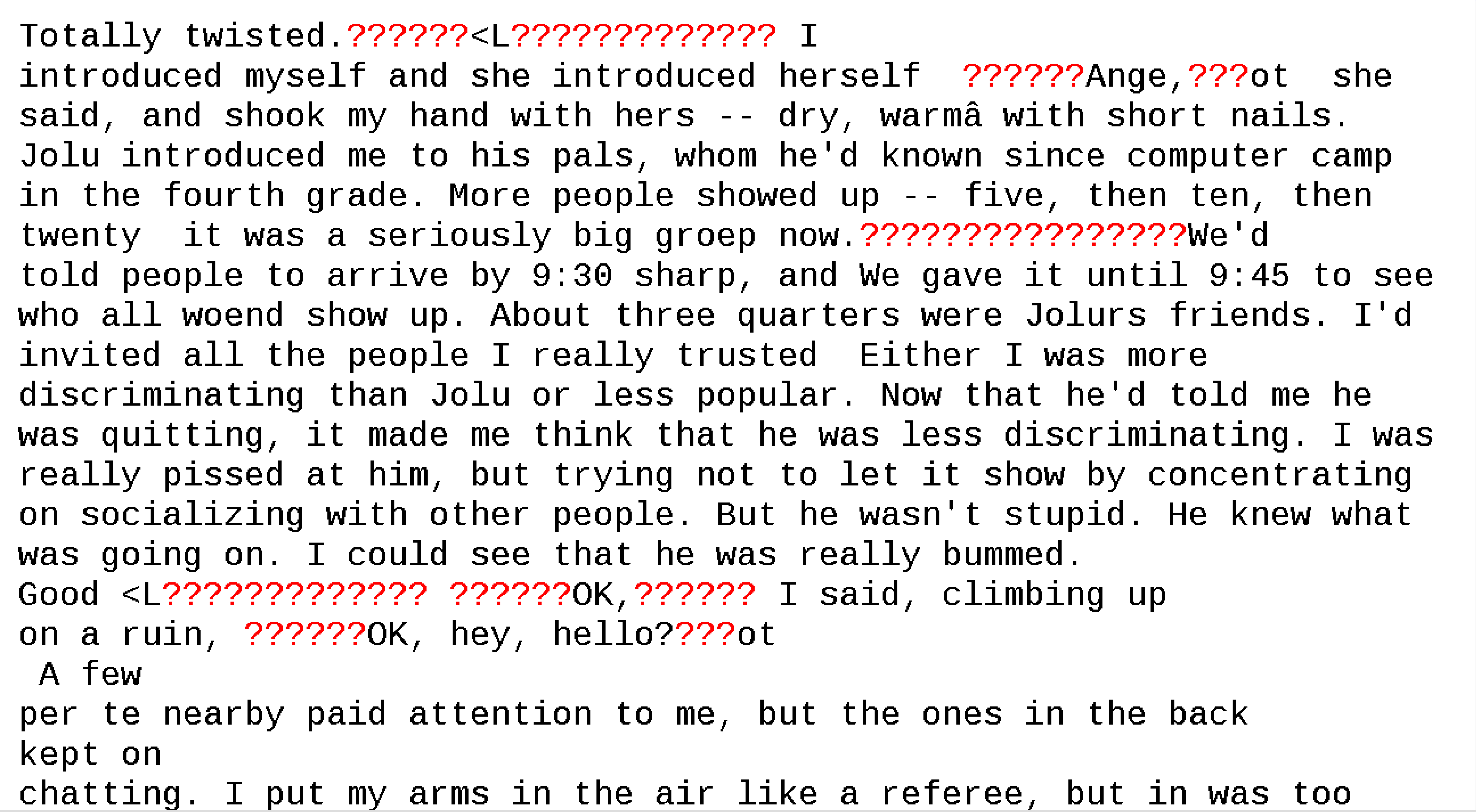
Example File

- HTML version of Cory Doctorow novel “Little Brother” (786,775 bytes)
  - Compressed using Info-Zip's zip version 3.0
  - First 1024 bytes of archive removed

Recovered Text Example

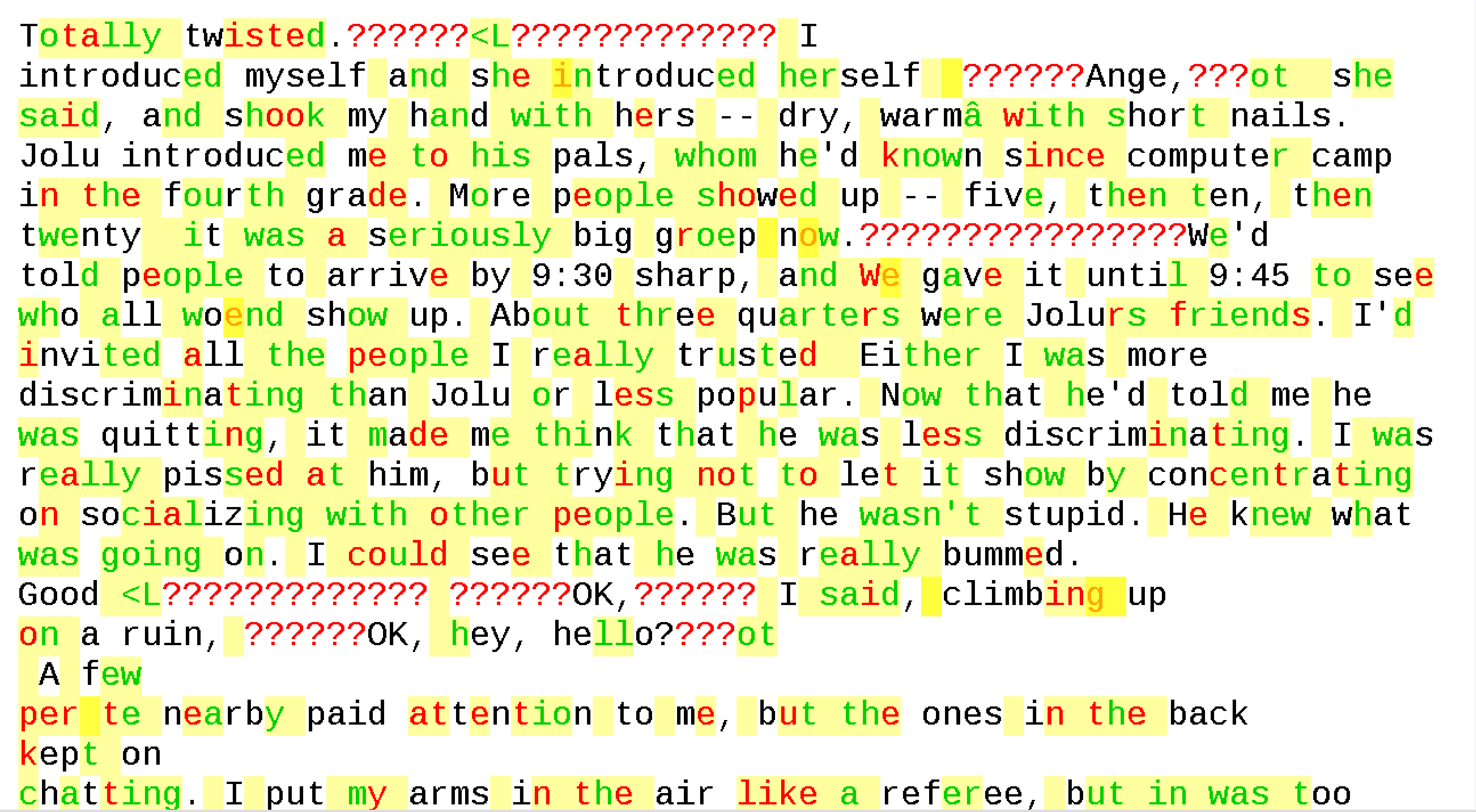
```
T???????tw?????.???????????????????????I
introduc?? myself?a???s?????troduc?? ???self????????Ange,???????s??
????, a???s?????my?h??d??????h?rs?--?dry,?warm????????hor??nails.
Jolu introduc?? m?????????pals,??????h?'d?????n?s?????compute??camp
i??????f??r???gra??.?M?re?p?????????w???up?--?fiv?,?t?????en,?t???
t??nty???t???????s?????????big?g???p?n??.????????????????W?'d
tol??p??????to?arriv??by?9:30?sharp,?a??????g?v??it?unti??9:45 ???se?
??o??ll??o????sh???up.?Ab???????e??qu???????w????Jolu??????????.?I'?
?nvi??? ?ll???? ???????I?r??????tr?s?e???Ei?????I???s?more
discrim??a???????an?Jolu??r?l????po?u?ar.?N?????at??e'd?tol??me?he
??? quitt???,?it??a???m?????n??t?at??e ??s?l????discrim??a????.?I???s
r??????pis???????him, b????ry??????? ???le??i? sh???b??con?????a????
o??so????iz?????????????????????.?B???he ???????stupid.?H??k????w?at
??? ??????o?.?I???????se??t?at??e ??s?r??????bumm?d.
Good???????????????????????OK,???????I?????,?climb????up
????a ruin,???????OK,??ey, he??o????????A?f??
???????n??rb??paid ??t?n???n?to?m?,?b?????? ones?i??????back??ep??on
?h?t????.?I?put????arms?i??????air ???????ref??ee,?b???????????oo
dark.?E???tual???I??i???n??????dea????turni???my LED??ey?h???????and
p??????? i???t??ach??f??????alke???i???urn,?t????at???. G?adual???
```

Reconstructed Text Example



```
Totally twisted.??????<L????????????? I
introduced myself and she introduced herself  ??????Ange,???ot  she
said, and shook my hand with hers -- dry, warmâ with short nails.
Jolu introduced me to his pals, whom he'd known since computer camp
in the fourth grade. More people showed up -- five, then ten, then
twenty  it was a seriously big groep now.????????????????We'd
told people to arrive by 9:30 sharp, and We gave it until 9:45 to see
who all woend show up. About three quarters were Jolurs friends. I'd
invited all the people I really trusted  Either I was more
discriminating than Jolu or less popular. Now that he'd told me he
was quitting, it made me think that he was less discriminating. I was
really pissed at him, but trying not to let it show by concentrating
on socializing with other people. But he wasn't stupid. He knew what
was going on. I could see that he was really bummed.
Good <L????????????? ??????OK,?????? I said, climbing up
on a ruin, ??????OK, hey, hello????ot
 A few
per te nearby paid attention to me, but the ones in the back
kept on
chatting. I put my arms in the air like a referee, but in was too
```

Reconstructed Text Example



Totally twisted.??????<L????????????? I
introduced myself and she introduced herself ??????Ange,???ot she
said, and shook my hand with hers -- dry, warmâ with short nails.
Jolu introduced me to his pals, whom he'd known since computer camp
in the fourth grade. More people showed up -- five, then ten, then
twenty it was a seriously big groep now.????????????????We'd
told people to arrive by 9:30 sharp, and We gave it until 9:45 to see
who all woend show up. About three quarters were Jolurs friends. I'd
invited all the people I really trusted Either I was more
discriminating than Jolu or less popular. Now that he'd told me he
was quitting, it made me think that he was less discriminating. I was
really pissed at him, but trying not to let it show by concentrating
on socializing with other people. But he wasn't stupid. He knew what
was going on. I could see that he was really bummed.
Good <L????????????? ??????OK,?????? I said, climbing up
on a ruin, ??????OK, hey, hello????ot
A few
per te nearby paid attention to me, but the ones in the back
kept on
chatting. I put my arms in the air like a referee, but in was too

Original Passage

```
Totally twisted.&quot;</FONT></P>  <p>I
introduced myself and she introduced herself. &quot;Ange,&quot; she
said, and shook my hand with hers -- dry, warm, with short nails.
Jolu introduced me to his pals, whom he'd known since computer camp
in the fourth grade. More people showed up -- five, then ten, then
twenty. It was a seriously big group now.</FONT></P>  <p>We'd
told people to arrive by 9:30 sharp, and we gave it until 9:45 to see
who all would show up. About three quarters were Jolu's friends. I'd
invited all the people I really trusted. Either I was more
discriminating than Jolu or less popular. Now that he'd told me he
was quitting, it made me think that he was less discriminating. I was
really pissed at him, but trying not to let it show by concentrating
on socializing with other people. But he wasn't stupid. He knew what
was going on. I could see that he was really bummed. Good.</FONT></P>

<p>&quot;OK,&quot;
I said, climbing up on a ruin, &quot;OK, hey, hello?&quot; A few
people nearby paid attention to me, but the ones in the back kept on
chatting. I put my arms in the air like a referee, but it was too
dark. Eventually I hit on the idea of turning my LED keychain on and
```

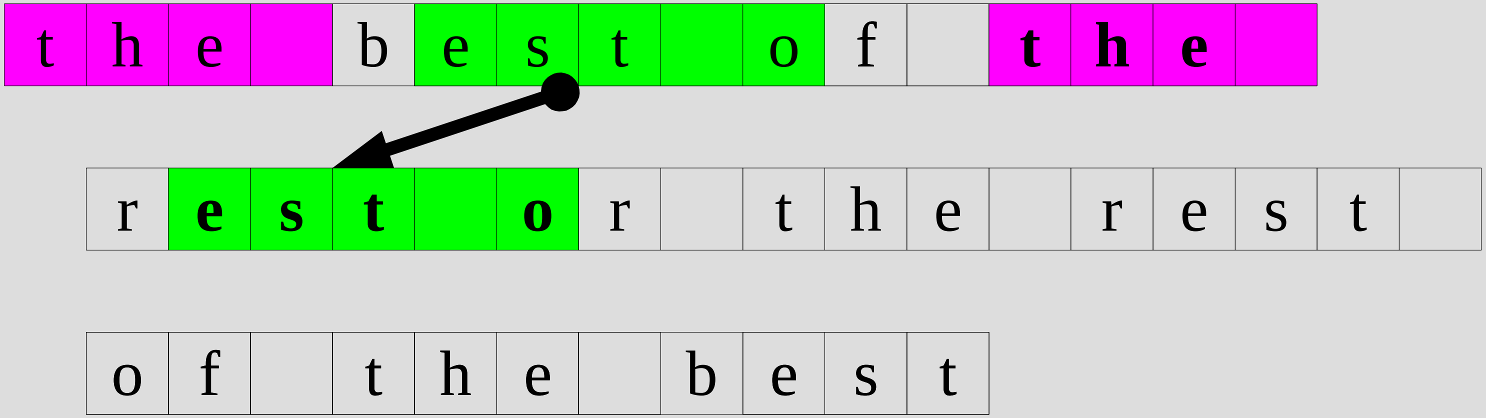
DEFLATE Compression

- By far the most common algorithm for ZIP files
- Two phases:
  - Replace repeated occurrences of multi-byte sequences within a 32 KB (optionally 64 KB) window with a reference to the previous occurrence
  - Apply Huffman coding to efficiently represent the mixed sequence of literal bytes and offset:length pairs
- Decompressor must track compressor's state
  - Missing the beginning of the bitstream prevents this

DEFLATE: Chaining Occurrences

t	h	e		b	e	s	t		o	f		t	h	e			
	r	e	s	t		o	r		t	h	e		r	e	s	t	
	o	f		t	h	e		b	e	s	t						

DEFLATE: Chaining Occurrences

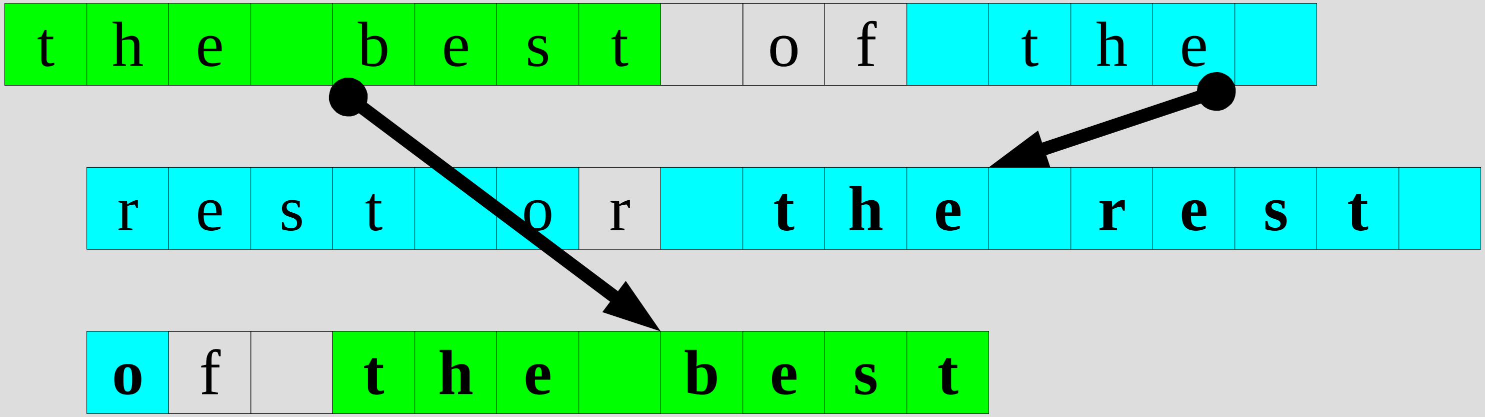


t h e b e s t o f **t h e**

r **e s t o** r t h e r e s t

o f t h e b e s t

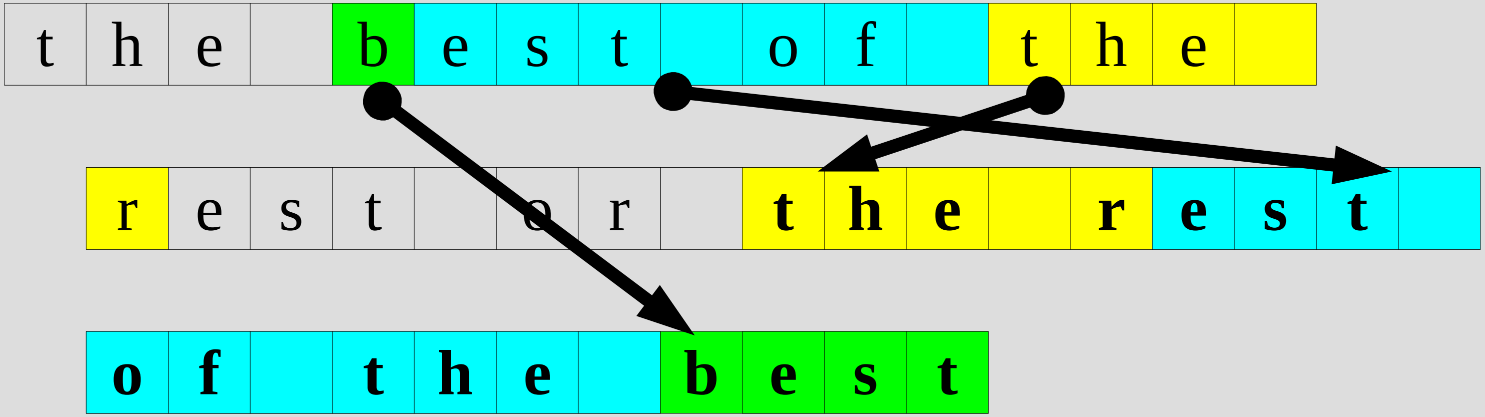
DEFLATE: Chaining Occurrences



the best of the

rest or **the rest**

of **the best**



the best of the

rest or **the rest**

of the best

DEFLATE: Chaining Occurrences

t	h	e		b	e	s	t		o	f		*12/4*	r	*12/5*

r	*12/11*	f		*36/8*

t	h	e		b	e	s	t		o	f		*12/4*	r	*12/5*

r	*12/6*	*24/11*	*36/4*

Recovering Compressor's State

- DEFLATE does not use adaptive Huffman coding, so the compressor breaks the stream into blocks, each of which may be
  - Uncompressed
  - Compressed with a predefined Huffman tree
  - Compressed with a tree transmitted in the stream
- Finding the start of a block gives us a known state for the Huffman compression
  - But not the contents of the back-reference window

Finding the Start of a Block

- Three-**BIT** header (block type and last-block flag)
- Header can appear at any bit position
- Need to scan at every bit position, testing whether a validly-decompressible block starts at that bit
  - Valid header and Huffman tree
  - No invalid bit sequences in data stream
- Park et al (2008) did exactly such a scan in a brute-force manner
  - reported speed of 7 **kilo**bytes per second

Efficiently Finding a Block Start

- Work from **end** of compressed stream
  - Provides a known end to each block
  - Eliminates half of the potential starting bits
- Do quick sanity checks before full decompression
  - is alphabet size legal?
  - is the Huffman tree of bit lengths legal?
  - if the Huffman tree passes muster, is there an end-of-data symbol at the end of the block?

Partial Decompression

- Once we have found the first intact block, we can decompress from that point forward
- However, references to text prior to that point will be unknown
- Initially, most bytes are unknown, but the proportion decreases as we progress
  - Bytes can remain unknown far beyond the 64 KB window if a reference is made to a sequence containing an unknown byte

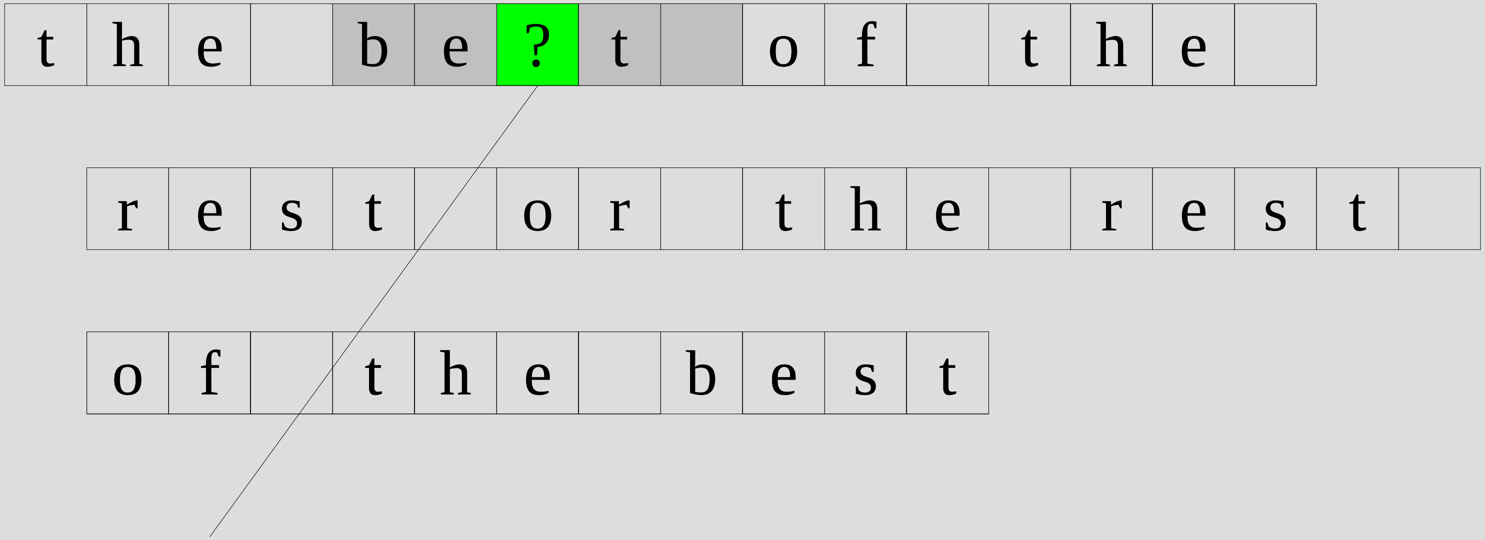
Recovered Text

```
T???????tw?????.??????????????????????I
introduc?? myself?a???s?????troduc?? ???self????????Ange,???????s??
????, a???s?????my?h??d??????h?rs?--?dry,?warm????????hor??nails.
Jolu introduc?? m?????????pals,??????h?'d?????n?s?????compute??camp
i??????f??r???gra??.?M?re?p?????????w???up?--?fiv?,?t?????en,?t???
t??nty???t???????s?????????big?g???p?n??.????????????????W?'d
tol??p??????to?arriv??by?9:30?sharp,?a??????g?v??it?unti??9:45 ???se?
??o??ll??o????sh???up.?Ab???????e??qu???????w????Jolu??????????.?I'?
?nvi??? ?ll???? ???????I?r??????tr?s?e???Ei?????I???s?more
discrim??a???????an?Jolu??r?l????po?u?ar.?N?????at??e'd?tol??me?he
??? quitt???,?it??a???m?????n??t?at??e ??s?l????discrim??a????.?I???s
r??????pis???????him, b????ry??????? ???le??i? sh???b??con?????a????
o??so????iz?????????????????????.?B???he ???????stupid.?H??k????w?at
??? ??????o?.?I???????se??t?at??e ??s?r??????bumm?d.
Good???????????????????????OK,???????I?????,?climb????up
????a ruin,???????OK,??ey, he??o????????A?f??
???????n??rb??paid ??t?n???n?to?m?,?b?????? ones?i??????back??ep??on
?h?t????.?I?put????arms?i??????air ???????ref??ee,?b???????????oo
dark.?E???tual???I??i???n??????dea????turni???my LED??ey?h???????and
p??????? i???t??ach??f??????alke???i???urn,?t????at???. G?adual???
```

Reconstructing Unknown Bytes

- Many of the unknown bytes have multiple occurrences
  - 75% of occurrences from copies of just 20% of the unknown bytes
- Many of those occurrences are the only unknown byte in a word
  - Can infer likely replacements
- Replacing some unknown bytes yields additional words from which we can infer replacements

Eliminating Impossible Values



the be?t of the

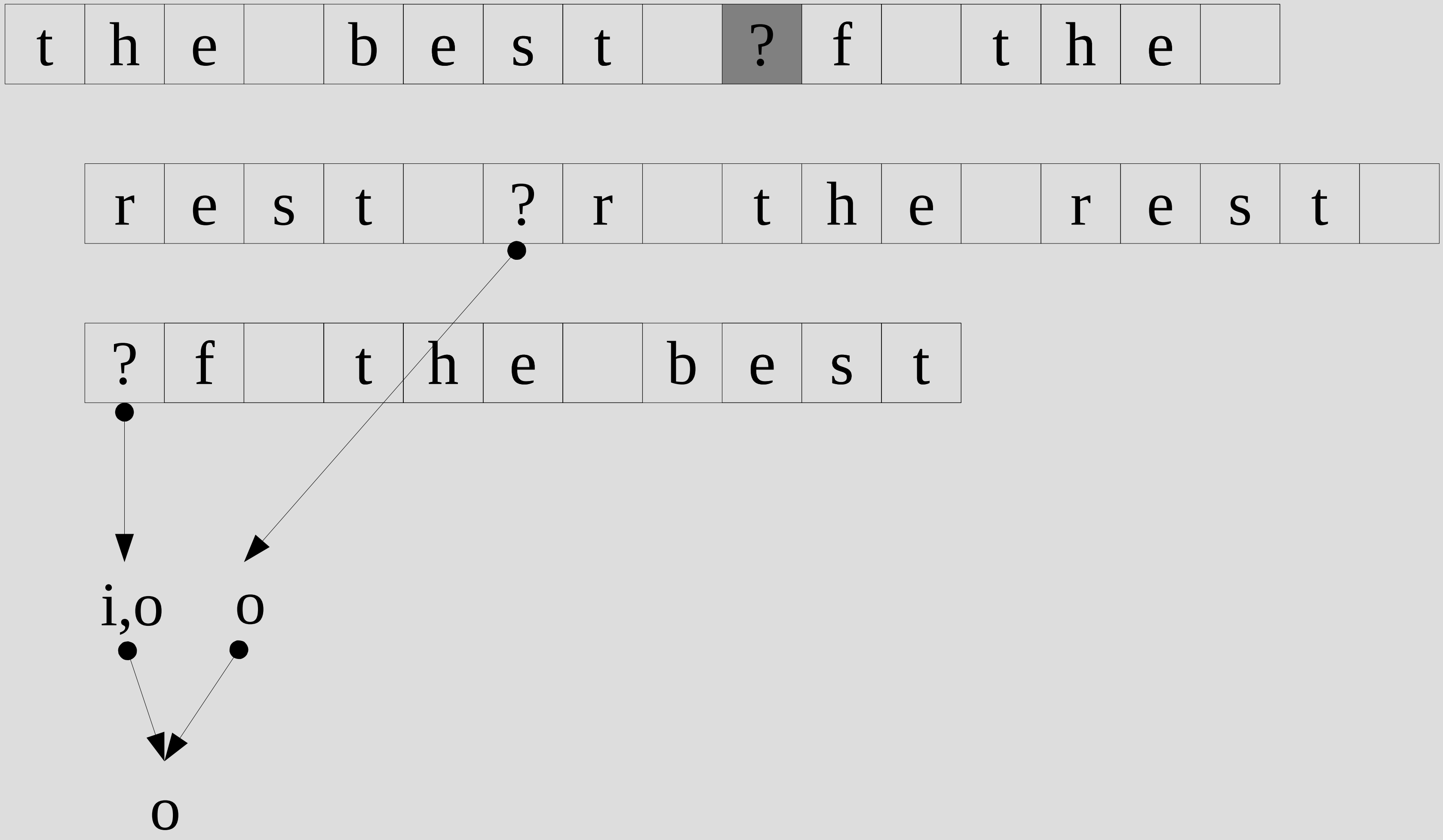
rest or the rest

of the best

Find all trigrams be? or e?t or “?t ” in training data.

Eliminate all values not supported by training data trigrams from consideration.

Inferring Unknown Bytes



the best ?f the

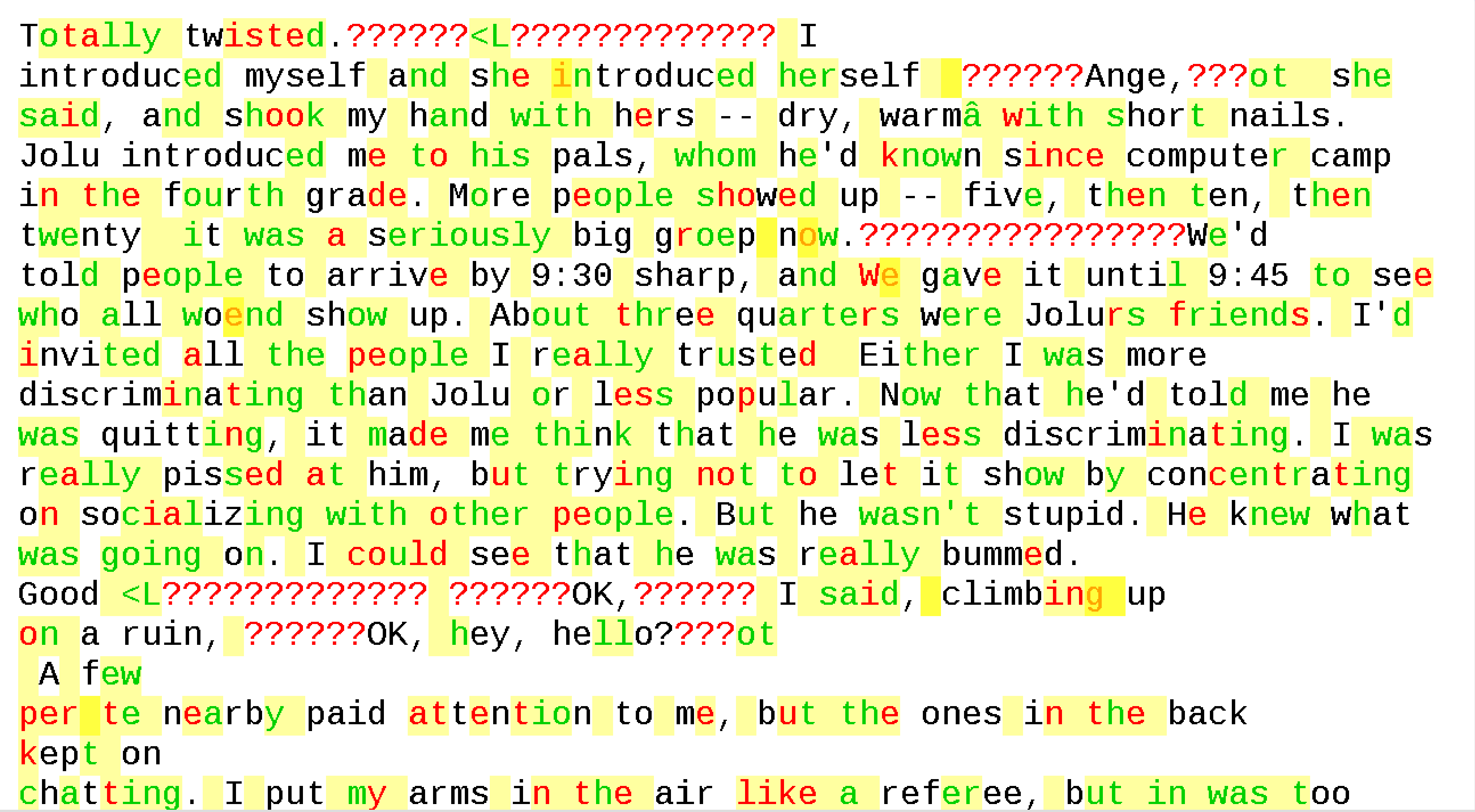
rest ?r the rest

?f the best

i,o o

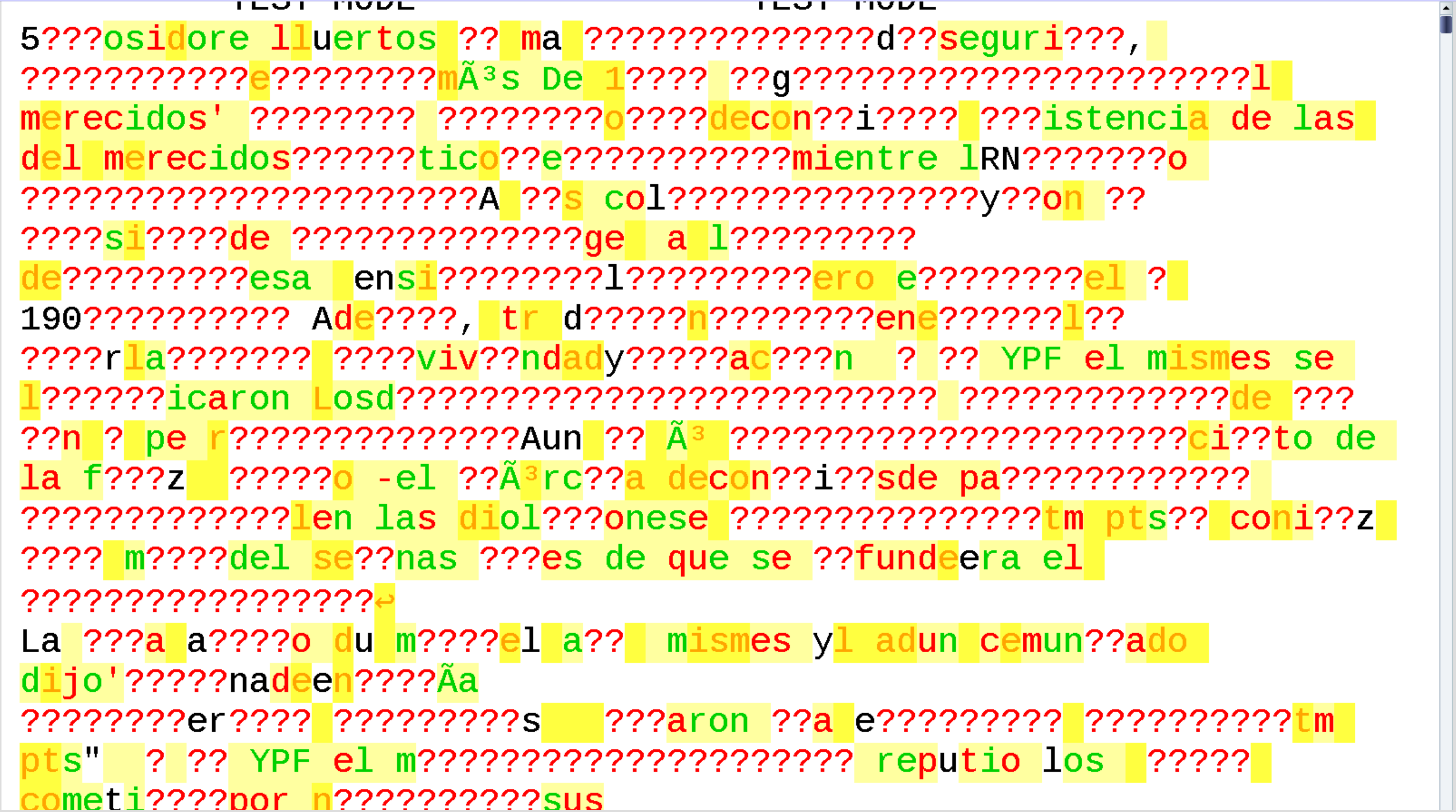
o

Reconstructed Text (English)



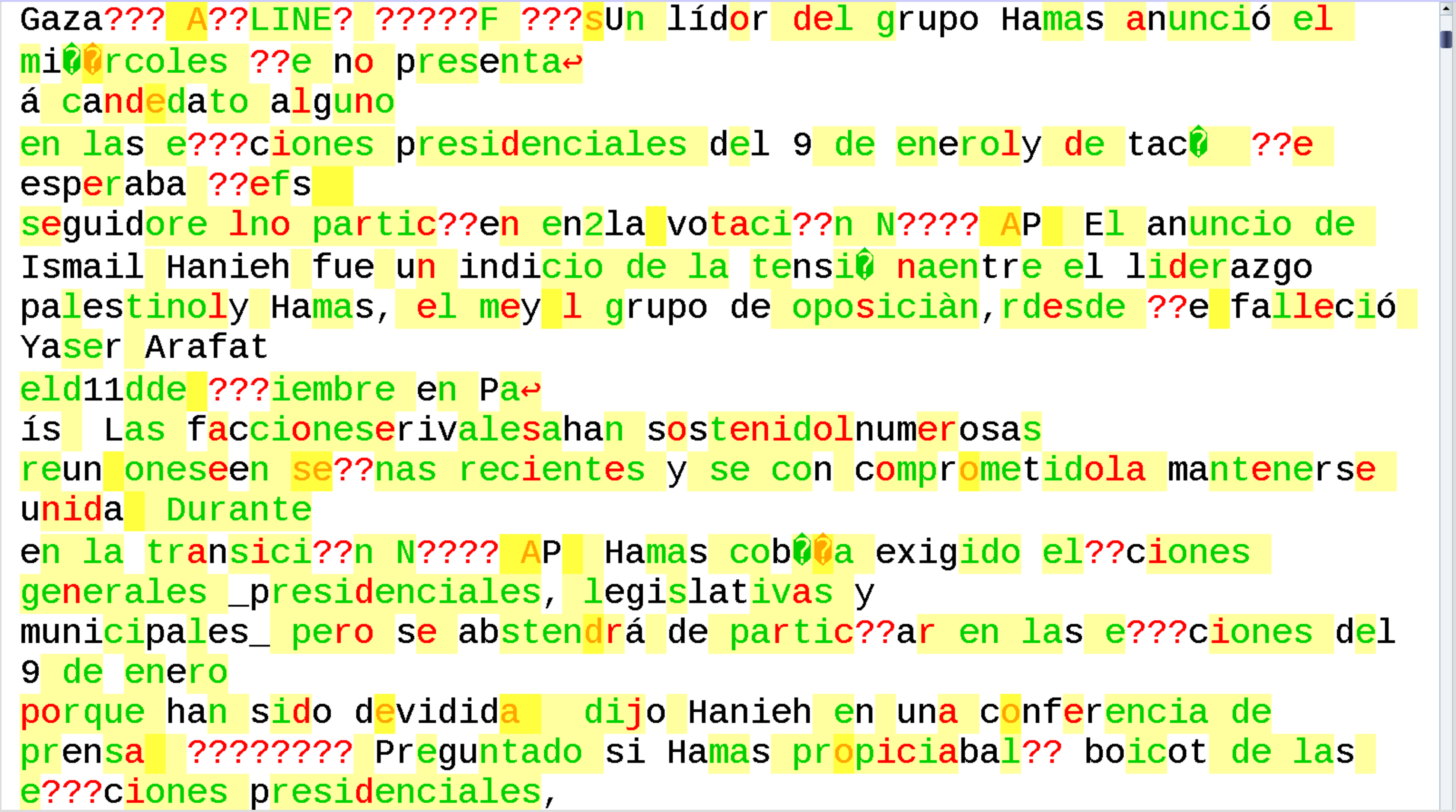
Totally twisted.??????<L????????????? I
introduced myself and she introduced herself ??????Ange,???ot she
said, and shook my hand with hers -- dry, warmâ with short nails.
Jolu introduced me to his pals, whom he'd known since computer camp
in the fourth grade. More people showed up -- five, then ten, then
twenty it was a seriously big groep now.????????????????We'd
told people to arrive by 9:30 sharp, and We gave it until 9:45 to see
who all woend show up. About three quarters were Jolurs friends. I'd
invited all the people I really trusted Either I was more
discriminating than Jolu or less popular. Now that he'd told me he
was quitting, it made me think that he was less discriminating. I was
really pissed at him, but trying not to let it show by concentrating
on socializing with other people. But he wasn't stupid. He knew what
was going on. I could see that he was really bummed.
Good <L????????????? ??????OK,?????? I said, climbing up
on a ruin, ??????OK, hey, hello????ot
 A few
per te nearby paid attention to me, but the ones in the back
kept on
chatting. I put my arms in the air like a referee, but in was too

Reconstructed Text (Spanish, start of recovery)



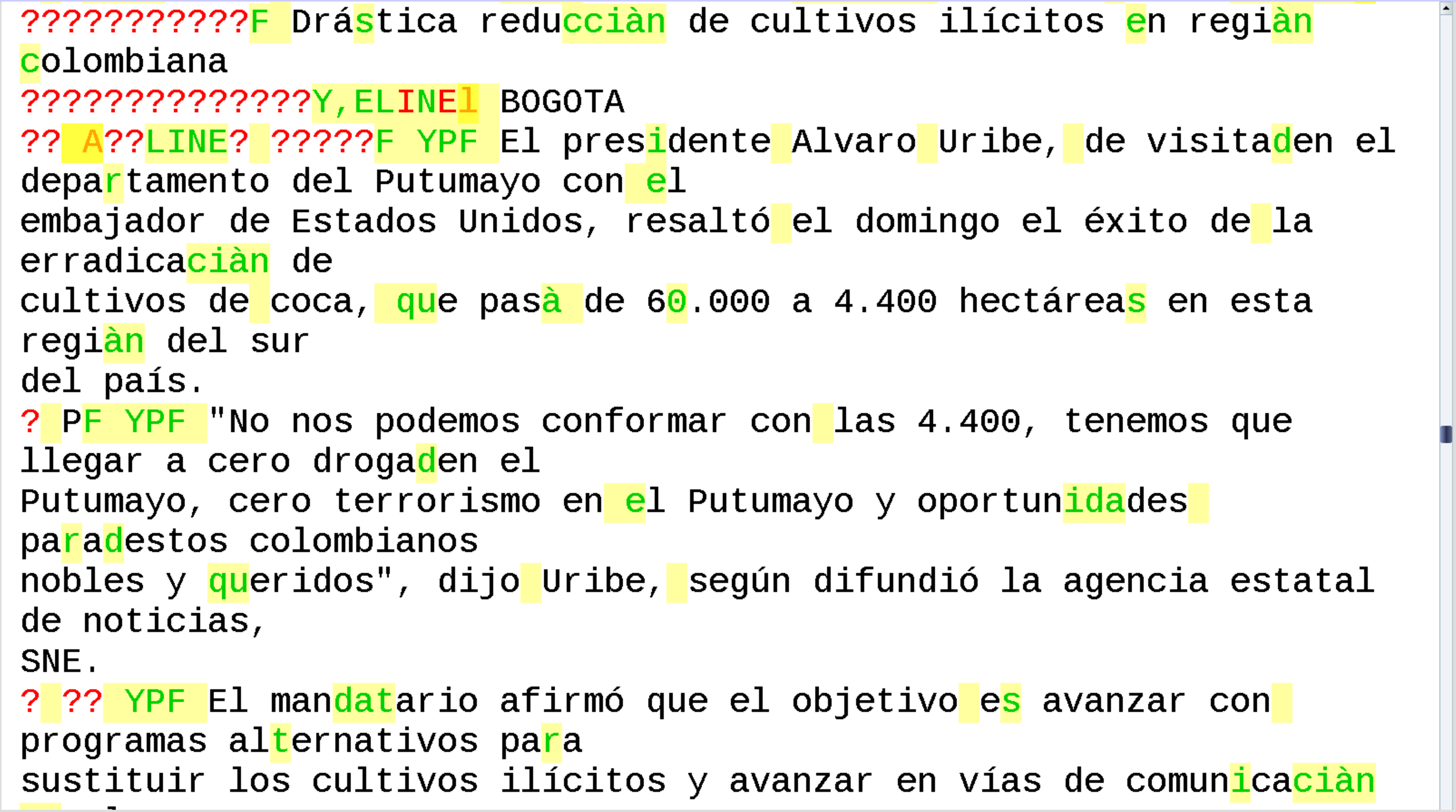
```
           TEST MODE                        TEST MODE
5???osidore lluertos ?? ma ??????????????d??seguri???, 
???????????e????????mÃ³s De 1???? ??g??????????????????????l 
merecidos' ???????? ????????o????decon??i???? ???istencia de las 
del merecidos??????tico??e???????????mientre lRN???????o
??????????????????????A ??s col???????????????y??on ??
????si????de ??????????????ge  a l?????????
de?????????esa  ensi????????l?????????ero e????????el ? 
190?????????? Ade????, tr d?????n????????ene??????l??
????rla??????? ????viv??ndady?????ac???n  ? ?? YPF el mismes se 
l??????icaron Losd?????????????????????????? ?????????????de ???
??n ? pe r??????????????Aun ?? Ã³ ??????????????????????ci??to de 
la f???z  ?????o -el ??Ã³rc??a decon??i??sde pa???????????? 
?????????????len las diol???onese ???????????????tm pts?? coni??z 
???? m????del se??nas ???es de que se ??fundeera el 
?????????????????↵
La ???a a????o du m????el a??  mismes yl adun cemun??ado 
dijo'?????nadeen????Ãa
????????er???? ?????????s   ???aron ??a e????????? ??????????tm 
pts"  ? ?? YPF el m????????????????????? reputio los  ????? 
cometi????por n??????????sus
```

Reconstructed Text (Spanish, a little further)



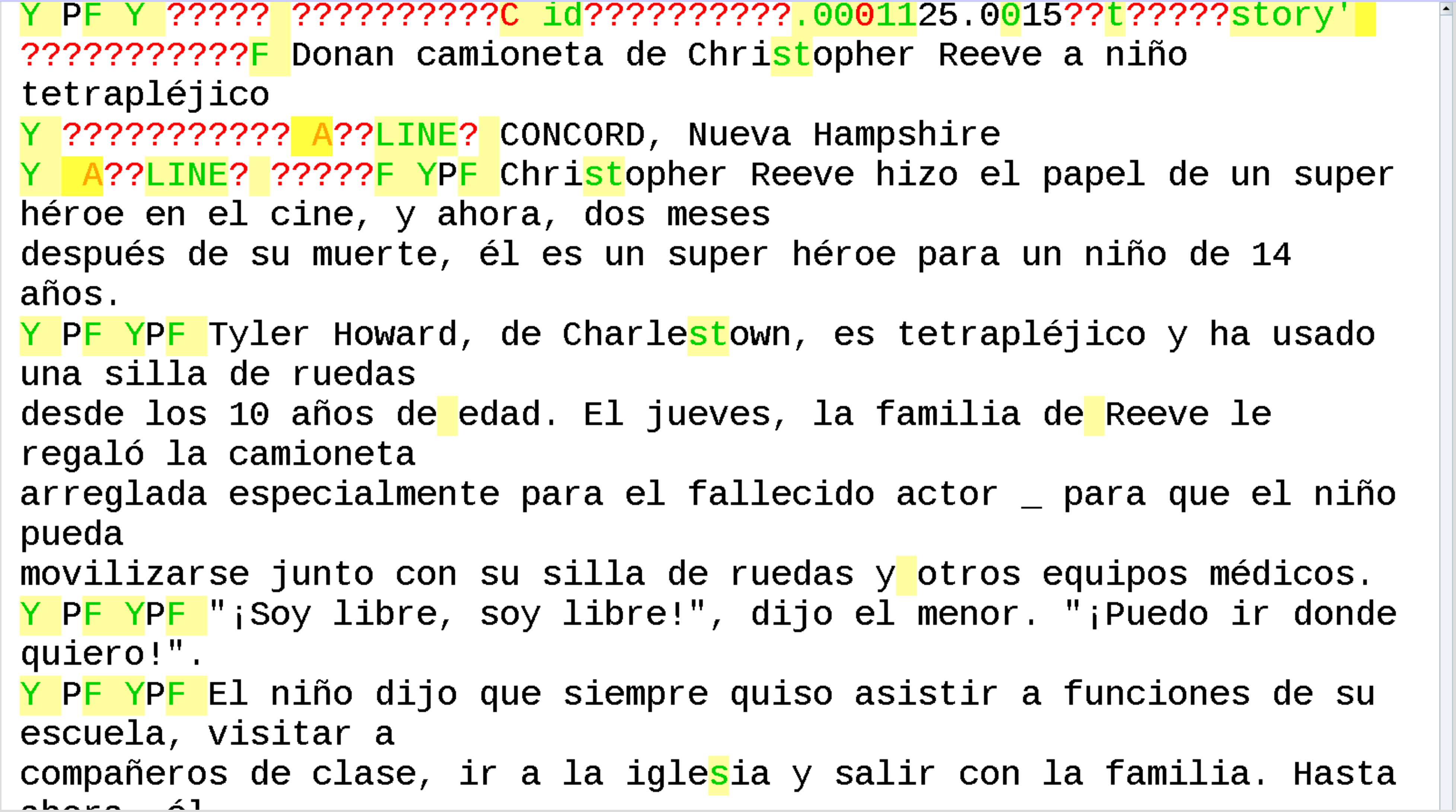
```
Gaza??? A??LINE? ?????F ???sUn lídor del grupo Hamas anunció el
mi��rcoles ??e no presenta↩
á candedato alguno
en las e???ciones presidenciales del 9 de eneroly de tac�  ??e
esperaba ??efs
seguidore lno partic??en en2la votaci??n N???? AP  El anuncio de
Ismail Hanieh fue un indicio de la tensi� naentre el liderazgo
palestinoly Hamas, el mey l grupo de oposiciàn,rdesde ??e falleció
Yaser Arafat
eld11dde ???iembre en Pa↩
ís  Las faccioneserivalesahan sostenidolnumerosas
reun oneseen se??nas recientes y se con comprometidola mantenerse
unida  Durante
en la transici??n N???? AP  Hamas cob��a exigido el??ciones
generales _presidenciales, legislativas y
municipales_ pero se abstendrá de partic??ar en las e???ciones del
9 de enero
porque han sido devidida   dijo Hanieh en una conferencia de
prensa  ???????? Preguntado si Hamas propiciabal?? boicot de las
e???ciones presidenciales,
```

Reconstructed Text (Spanish, half-way)



```
???????????F Drástica reducciàn de cultivos ilícitos en regiàn
colombiana
??????????????Y,ELINEl BOGOTA
?? A??LINE? ?????F YPF El presidente Alvaro Uribe, de visitaden el
departamento del Putumayo con el
embajador de Estados Unidos, resaltó el domingo el éxito de la
erradicaciàn de
cultivos de coca, que pasà de 60.000 a 4.400 hectáreas en esta
regiàn del sur
del país.
? PF YPF "No nos podemos conformar con las 4.400, tenemos que
llegar a cero drogaden el
Putumayo, cero terrorismo en el Putumayo y oportunidades
paradestos colombianos
nobles y queridos", dijo Uribe, según difundió la agencia estatal
de noticias,
SNE.
? ?? YPF El mandatario afirmó que el objetivo es avanzar con
programas alternativos para
sustituir los cultivos ilícitos y avanzar en vías de comunicaciàn
```

Reconstructed Text (Spanish, end of file)



```
Y PF Y ????? ??????????C id??????????.0001125.0015??t?????story'
???????????F Donan camioneta de Christopher Reeve a niño
tetrapléjico
Y ??????????? A??LINE? CONCORD, Nueva Hampshire
Y  A??LINE? ?????F YPF Christopher Reeve hizo el papel de un super
héroe en el cine, y ahora, dos meses
después de su muerte, él es un super héroe para un niño de 14
años.
Y PF YPF Tyler Howard, de Charlestown, es tetrapléjico y ha usado
una silla de ruedas
desde los 10 años de edad. El jueves, la familia de Reeve le
regaló la camioneta
arreglada especialmente para el fallecido actor _ para que el niño
pueda
movilizarse junto con su silla de ruedas y otros equipos médicos.
Y PF YPF "¡Soy libre, soy libre!", dijo el menor. "¡Puedo ir donde
quiero!".
Y PF YPF El niño dijo que siempre quiso asistir a funciones de su
escuela, visitar a
compañeros de clase, ir a la iglesia y salir con la familia. Hasta
ahora, él
```

Limitations to Reconstruction

- Word-based
  - Will not work well with languages that don't use spaces
  - Current code can't handle multi-byte non-word characters
- Needs an appropriate language model
  - Differences between training data and the file being reconstructed degrade accuracy
    - Mitigated by adding recovered literal text to model
  - Currently must supply the correct model manually

Efficacy (1)

- Run in test mode, simulating a missing first byte for every archive member
- On ZipRec v0.9 source code (286 files, 3.8 MB)
  - 21 files consist of multiple packets
  - 97,053 literal bytes, 654,700 total bytes recoverable
- On a collection of downloaded zip archives (79 archives, 148 MB; containing 8310 files totalling 336 MB)
  - 859 files consist of multiple packets
  - 134 MB literal bytes, 199 MB total recoverable

Efficacy (2)

- On disk image UAE10-009 from Real Data Corpus:
  - Detects
    - 10,478 local file header signatures
    - 11,725 central directory entries
    - 550 end of central directory records
  - Extracts
    - 6922 complete files (5309 short and stored uncompressed)
    - 446 partial files
    - Total 78 MB, of which 77 MB literal bytes

Speed

- On the novel we have been using as an example:
  - unzip (intact file): 30ms
  - ZipRec recover: 290ms
  - ZipRec reconstruct: 58,000ms – 69,000ms
- On the ZipRec source code:
  - unzip (intact file): 105ms
  - ZipRec recover: 795ms
  - ZipRec reconstruct: 24,000ms
- Scanning disk image from Real Data Corpus:
  - about 2 minutes per gigabyte, including recovery

Future Work

- Improved recovery
  - attempt to decompress the initial partial block using information from a first-pass reconstruction
- Improved reconstruction
  - automatic language identification to select proper model
  - higher-order language models
- GUI to manually fix reconstruction

ZipRec is Open Source

- Get it now:
  - http://ziprec.sourceforge.net/
- Download includes C++ source code, sample language models, and 64-bit Linux executable

Questions?

Search Statistics

```
Found 0 local and 1 central file headers
Uncompressed packets:
  268418 candidates
  0 valid
Fixed-Huffman packets:
  272549 candidates
  0 considered
  0 valid
Dynamic-Huffman packets:
  273632 candidates
  233670 with valid alphabet sizes
    154464 had invalid bit-length tree
    79061 had invalid bit lengths
  130 with valid Huffman tree
  4 with valid EOD marker
  4 valid
```

When to use ZipRec

- When a standard unzip program fails
  - ZipRec will work on intact archives, but is 8-10x slower
- When missing parts of a split archive
  - Concatenate available parts in order and apply ZipRec
- When a file may contain multiple archives
  - Standard programs may only see some of the files

What to Do if ZipRec Fails

- Check that the file is a ZIP archive or contains one
  - ZIPX extra compression types only partially supported
    - Uncorrupted BZIP2 and WavPack blocks can be extracted
- If using a file carver, try running ZipRec on the original image
  - Could take a long time, but ZipRec will handle multi-terabyte files on 64-bit systems
- Is your file fragment big enough?
  - Must contain either the start or end of a compressed file, plus the adjacent header

```
1743024 dynamic-Huffman packet candidates
1486010 with valid alphabet sizes
   986690 invalid bit-length trees, 498299 invalid bit lengths
869 with valid Huffman tree
34 with valid EOD marker, of which 32 valid

962946 total unknown bytes (354161 not reconstructed)
18037 distinct words with unknown bytes processed
1444 of 6597 co-indexed classes replaced
492759 of 608786 reconstructed bytes correct (80.9%)

  0.01s scanning for members
  1.30s searching for packets
  0.20s inflating
  0.31s extracting reference file
221.32s reconstructing
    29.56s collecting trigram constraints
   188.47s scoring candidates
```

General Applicability

- Will this approach work with other compressors?
  - Reconstruction can be applied to any Lempel-Ziv type sequence of mixed literals and back-references
  - Getting that L-Z sequence may be more difficult with other compressors
    - e.g. LZMA uses adaptive entropy coding and does not have restart points
  - Other programs using DEFLATE simply need the appropriate signatures for start and end
    - ZipRec recognizes the ALZip signatures as well as PKZip