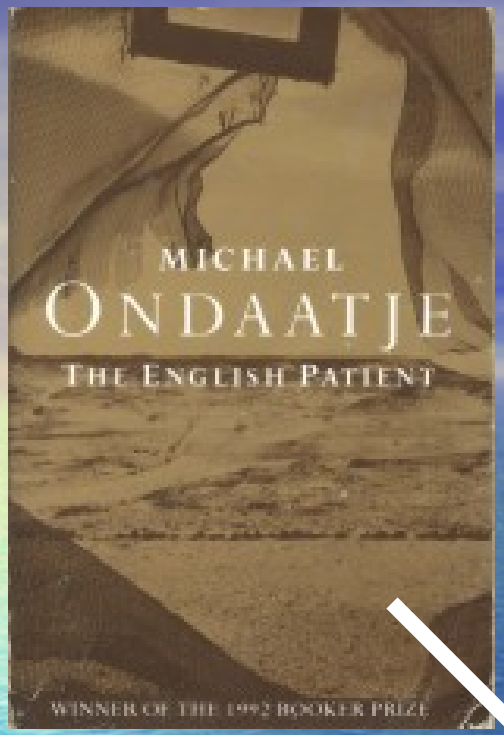
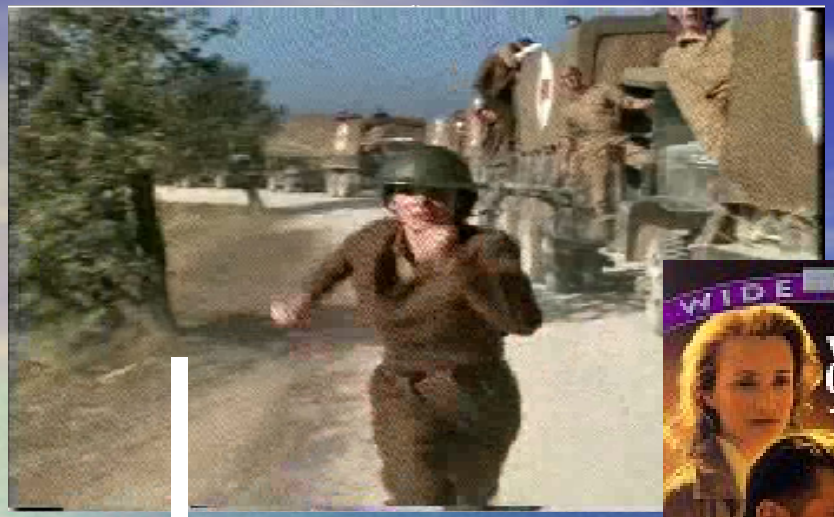
Analysing Collateral Text Corpora

Dr Andrew Salway
Department of Computing, University of Surrey

11th April 2005


MICHAEL
ONDAATJE
THE ENGLISH PATIENT
WINNER OF THE 1992 BOOKER PRIZE


Based on the Novel by Michael Ondaatje
The English Patient
A Screenplay
Anthony Minghella
Now a Major Motion Picture from



AUDIO DESCRIPTION

[03.03] The pilot of the open plane is wearing goggles and a leather flying helmet

[03:22] Clearly visible against a cloudless sky, the plane flies on over the rolling sand hills.

[03:27] German gunners spot the aircraft…

PLOT SUMMARY

The moving story of an English mapmaker and his dying memories of the romance that tragically alters his life. Burned horribly in a fiery crash after being shot down while crossing the Sahara Desert during WWII, he is tended to by a Canadian nurse …

Analysing Collateral Texts for Films

Why?

- **To look for and describe the use of special language**
- **To investigate narrative structures**
- To enable novel retrieval and browsing of video data

How?

- **Corpora Analysis: statistical measures of linguistic variance and local grammar fragments**
  - Audio description
  - Screenplays
  - Plot summaries

Special Language
or, Language for Special Purpose (LSP)

- "language looks and behaves the way it does because of the work it does for us" (Brandt 1986)

  → distinct language registers…

- Seven approaches to LSP (Hoffman 1984), inc.:
  - Restricted field of discourse → restricted vocabularly / terminological studies
  - Functional explanations / map between communicative need and prevalence of certain lexicogrammatical constructions, cf. 'functional tenor'

Corpus Linguistics

- Can characterise a language register, including a special language, by a combination of statistically significant linguistic features which appear in a corpus of special language texts compared with a general language sample

(Biber, Conrad and Reppen 1998)

Local Grammar

- Concentration on word classes and substitution in local contexts, to avoid 'overgeneralization' whilst accounting for all possible sentences within a corpus, Gross (1997)
- Finite-state automata approach has resonance with work in Information Extraction

Narrative

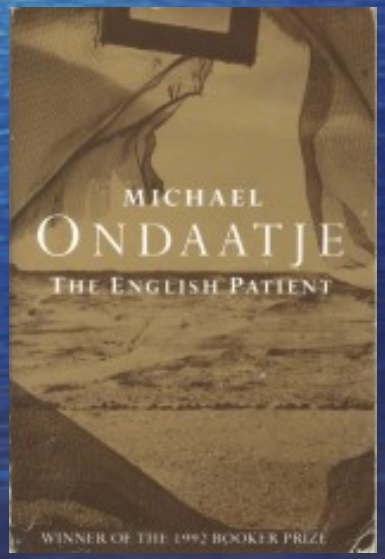
- The study of narrative explores what the media-independent features of stories are, how different kinds of media can convey (the same) stories, and how stories are understood (Ryan 2004). Can be actual or fictional stories told in any combinations of media
- For some, narrative abilities considered both as a mode of thought and of discourse are fundamental to intelligence (Bruner 1991, Schank 1990)
- Narrative is a multi-faceted phenomenon studied by philosophers, literature and film scholars, linguists, cognitive scientists and computer scientists (Herman 2002)

Narrative

- Distinction made between ***story*** and ***discourse*** (Chatman 1978)

Story

Discourse



Discourse

Narrative

- Narrative involves chains of events in cause-effect relationships occurring in space and time, where the agents of cause-effect are characters with goals, beliefs and emotions (Bordwell and Thompson 1997).

- **BUT…** "***More than reconstructed timelines and inventories of existents***, … interpreters are called upon to live out complex blends of cognitive and imaginative response, encompassing sympathy, the drawing of causal inference, identification, evaluation, suspense, and so on" (Herman 2002)

Hypothesis

- If the 'function' of audio description, film screenplays and plot summaries is to tell/retell the story of films, then maybe the idiosyncracies of their special languages will reflect important aspects of (media-independent) narrative structures?

Corpus Analysis Methods

- Automated corpus analysis to identify idiosyncratic lexicogrammatical features, or **linguistic variance** (Ahmad and Rogers 2001) – statistically-based tests for differences of lexical, morphological and collocational features (Ahmad and Rogers 2001)
- From concordances → collocations → local grammars (Traboulsi, Cheng and Ahmad 2004)
- Using:
  - **System Quirk** - a language engineering workbench developed by researchers at the University of Surrey, and available as freeware from: www.computing.surrey.ac.uk/SystemQ/
  - **Unitex** - a corpus processing system, based on automata-oriented technology developed at Laboratoire d'Automatique Documentaire et Linguistique). Available from: http://www-igm.univ-mlv.fr/~unitex/

Audio Description

Audio Description Script

[11.43] Hanna passes Jan some banknotes.

[11.55] Laughing, Jan falls back into her seat as the jeep overtakes the line of the lorries.

[12.01] An explosion on the road ahead.

[12.08] The jeep has hit a mine.

[12.09] Hanna jumps from the lorry.

[12.20] Desperately she runs towards the mangled jeep.

[12.27] Soldiers try to stop her.

[12.31] She struggles with the soldier who grabs hold of her firmly.

[12.35] He lifts her bodily from the ground, holding her tightly in his arms.

(NB. Some 'cue' information removed)

Audio Description

- In between existing dialogue a describer gives important information about on-screen scenes and events, and about characters' actions, appearance, gestures and expressions.
- Enhances the enjoyment of most kinds of films and television programs for visually impaired viewers. Provided with some digital television broadcasts and with films in some cinemas and on VHS/DVD releases. Legislation / regulation mean audio description is increasingly available around the world.
- In the UK audio description is prepared by professionals who follow established guidelines. Guidelines suggest use of present tense and simple sentences, and the avoidance of ambiguous pronominal references.
- 8000-word audio description for a 2-hour film may take 60 person hours, with many viewings and more than one describer: however a 30 minute soap opera (full of dialogue and familiar scenes and characters) may take only 90 minutes to describe.

➔In effect the part of the story told by the moving image is retold in words.

Audio Description: production

- Normally scripted before it is recorded - 'written to be spoken' and includes time-codes to indicate when each utterance is to be spoken.
- Describers ensure audio description interacts with existing dialogue and sound effects
- They are constrained by the time available for description
- They have to be careful to strike the right balance between frustrating the audience with insufficient information, and patronizing them by spelling out obvious inferences – just enough information to follow the story?

Audio Description Corpus

- 420,000 words of audio description scripts from three producers of audio description in the UK: ITFC, Royal National Institute of the Blind (RNIB) and the BBC Audio Description Unit
- The majority of the corpus is audio description of feature films – 57 films in the 9 categories
- The remainder of the corpus consists of audio description for television programmes including documentaries, drama series and soap operas

Audio Description: Results

- Unusually high occurrence of open-class words in the top 100
- Unusually frequent, or 'weird', words…
  - referring to material processes
  - relating to characters' emotions
  - relating to temporal information
- Some seem to be part of local grammar fragments, e.g. *look*, *turns*, *smiles*, *door*, *room*

"Unusually" frequent…

Verbs referring to material processes

- 75% of commonly occurring verbs refer to material processes, for example:
  - *turn, sit, open, go, walk*

"Unusually" frequent…

Indicators of characters' emotions

- Adverbs - *anxiously, happily, nervously, desperately, sadly*
- Nouns - *hope, shock, relief, satisfaction*
- Adjectives - *worried, terrified, proud, nervous*

"Unusually" frequent…

Indicators of temporal information

NB. This analysis was based on 70,856 words of audio description scripts for 12 movies

- 200+ occurrences of aspectual verbs: *stop, start, begin* and *finish*
- 350 occurrences of *as* in temporal sense (often suggests a connection between events); 37 occurrences of *while* (only conveys simultaneity)
- 49 occurrences of *when / until* locate start / end of events in relation to other events and states
- 141 occurrences of *again* generally indicates a second instance of an event within a scene
- Some frequent adverbs may denote event duration, e.g. *still, slowly, quickly*

"Unusually" frequent…

Indicators of temporal information

NB. This analysis was based on 70,856 words of audio description scripts for 12 movies

- 173 occurrences of *then* – redundant for ordering of events, but seems to suggest completion of first event and sometimes the meeting of events
- 40 occurrences of *now* – adds no temporal information but seems to indicate a contrast between two events
- *after / before* are relatively infrequent and add nothing to text order
- Times of day like *night, morning, evening, dusk, dawn* are sometimes also used to introduce scenes; also, *later* (32 occurrences). Months, seasons, festival days and specific years were infrequent.

Other temporal information

- Timecodes and Text order
- Tense / Aspect:
  - present proliferates (even for events that are in future)
  - continuous used 'interchangeably' with simple present
  - occasional present perfect (when only end state of event is depicted, or to identify unnamed characters)
  - non-finite verbs with subordinate clauses indicate inclusion of event
- Inherent aspect of events (lexical aspect), e.g. punctual/durative, telic/atelic (Comrie 1976); entailment relations for verbs, e.g. co-extensive, proper inclusion, backward presupposition, (Fellbaum 1998)
- Time of day / historical period conveyed by costumes, props and lighting – all referred to in audio description

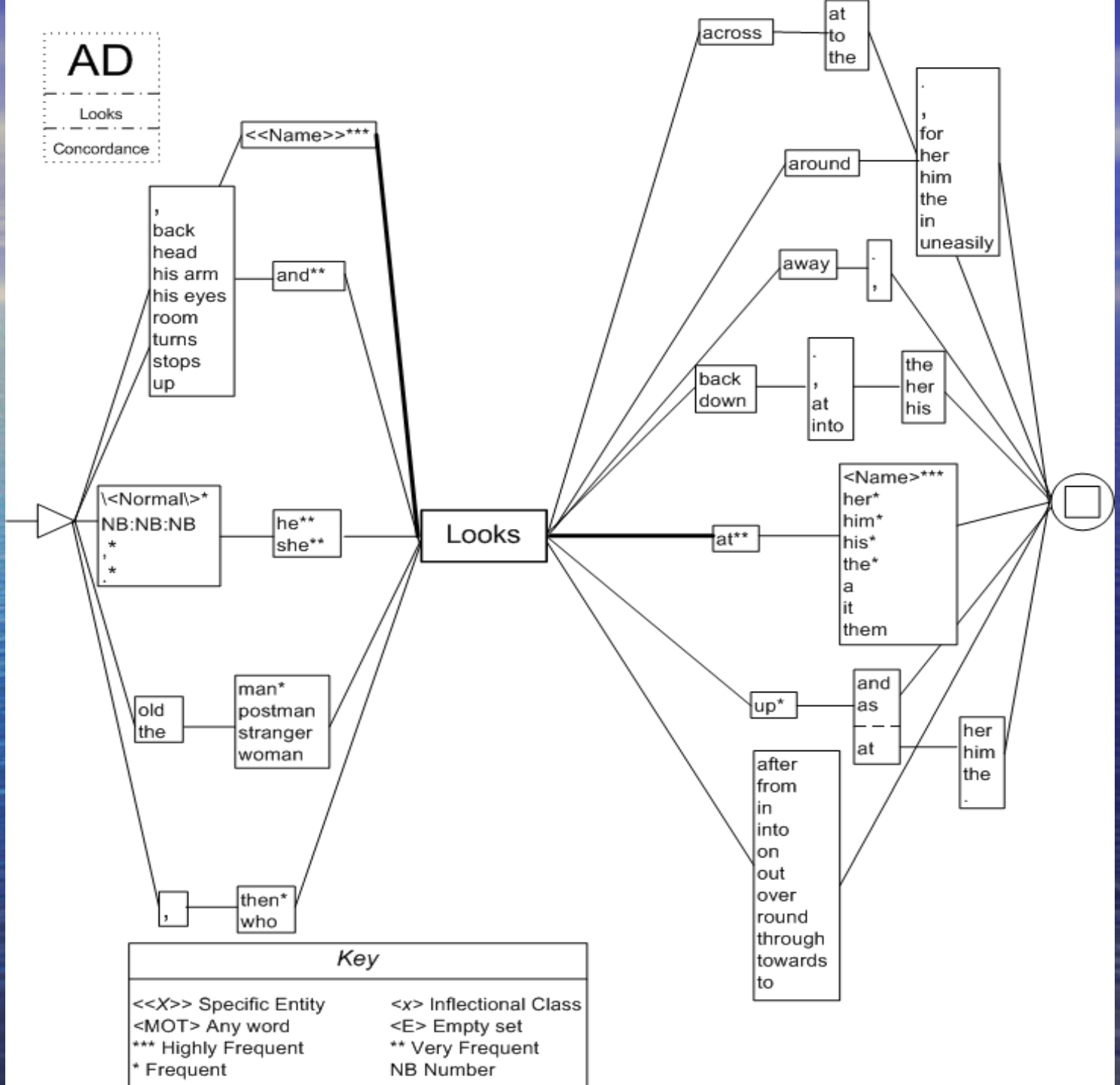
Local Grammar Fragments?

collocations

→ **common phrases**: 'look* at', 'turn* to', 'smiles at', 'open/close door', 'enter/leave room'

→ **local grammar fragments**

Maybe the high frequency of these phrases indicates important elements in filmic narratives?


AD
Looks
Concordance
Looks
<<Name>>***
,
back
head
his arm
his eyes
room
turns
stops
up
and**
\<Normal\>*
NB:NB:NB
.*
,*
.
he**
she**
old
the
man*
postman
stranger
woman
,
then*
who
across
at
to
the
around
,
for
her
him
the
in
uneasily
away
.
,
back
down
.
,
at
into
the
her
his
at**
<Name>***
her*
him*
his*
the*
a
it
them
up*
and
as
at
her
him
the
.
after
from
in
into
on
out
over
round
through
towards
to
Key
<<X>> Specific Entity
<MOT> Any word
*** Highly Frequent
* Frequent
<x> Inflectional Class
<E> Empty set
** Very Frequent
NB Number

Film Screenplay

7*. INT. RED CROSS TRUCK. DAY. A young CANADIAN NURSE, HANA, sits in a truck full of patients. Hana pays special care to the PATIENT lying in the stretcher alongside her. This is the PILOT - now known as THE ENGLISH PATIENT. A web of scars covers the Patient's face and body. They have the quality of a livid tattoo, magenta and green-black. The hair has largely gone and the effect is curious, lassoing his features, the strong nose, the eyes liquid. It's a warrior's face. But he has no physical strength. He coughs violently as the trucks shudders along the road.

8*. EXT. ITALIAN HILL ROAD. DAY. A JEEP pulls out of the line and approaches the Red Cross truck containing Hana and the Patient. The horn blows and Hana looks out to see it contains her best friend, JAN. TWO YOUNG SOLDIERS sit up front, one driving, both grinning. Jan signals for Hana's attention.

JAN There's meant to be lace in the next village - the boys are taking me.

HANA I'm not sewing anything else.

JAN (mischievously) You don't have any money, do you? Just in case there's silk.

HANA No!

JAN Hana, I know you do!

Film Screenplay Corpus

- 1,930,000 words for 71 films.
- About 3-4 times longer than audio description, i.e. much more detail, but a similar-looking language (except dialogue, and some cues like ‘INT’ and ‘EXT’)

Corpus	Open Class words (in first 100)	Total words
Film Screenplay	out, int, can, looks, like, day, ext, know, night, room, see, door, man, will, right, look, go, head, turns, eyes, hand, time, face, going, cut, come, think, car	**29**
Audio Description	out, looks, door, man, turns, head, eyes, hand, face, room, takes, walks, car, sits, hands, white, stands, tom, men, open, john, side, pulls, smiles, stares, goes, look, round, puts, steps, front, watches, water, opens, table, black, window, runs, stops, woman, bed	**41**
Common to both	**out, looks, door, room, man, look, head, turns, hand, eyes, face, car**	**12**

Plot Summary

Beginning in the 1930's, "The English Patient" tells the story of Count Almasy who is a Hungarian map maker employed by the Royal Geographical Society to chart the vast expanses of the Sahara Desert along with several other prominent explorers. As World War II unfolds, Almasy enters into a world of love, betrayal, and politics that is later revealed in a series of flashbacks while Almasy is on his death bed after being horribly burned in a plane crash.

Plot Summary Corpus

- 15,500 words for 114 films
- Compared with audio description and screenplay, the plot summaries seem to refer to 'higher-level' events
  - i.e. different frequent verbs…

Frequent verbs

Audio description: 56 films, 350,000 words, 41 OCW/100

Material:	open, walk, run, step, hold, close, go, wear, fall, lift, stand, throw, carry, kiss, sit, lead, get, give, cross, join, make, jump
Relational:	be
Mental:	watch, see
Behavioural:	smile, stare, look, glance, nod

Plot summaries: 114 films, 15,500 words, 28 OCW/100

Material:	do, get, find, take, kill, help, go, become, die, give, come, escape, make, murder, try, turn, change, follow, lose, need, run
Relational:	be, have
Verbal:	tell
Mental:	love, want, know, plan, decide, seem

Summary

- A number of idiosyncratic lexicogrammatical features in the three corpora suggest LSPs
- Possible 'functional' explanations
- Some insights into possible media-independent narrative structures (more tangible via language than moving images?)

Acknowledgements

- **Eleftheria Tomadaki**: analysis of audio description and plot summaries

- **Andrew Vassiliou**: analysis of audio description and screenplays