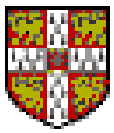


UNIVERSITY OF CAMBRIDGE

Retrieving Hierarchical Text Structure from Typeset Scientific Articles

Bill Hollingsworth

Ian Lewin (presenter)

Dan Tidhar

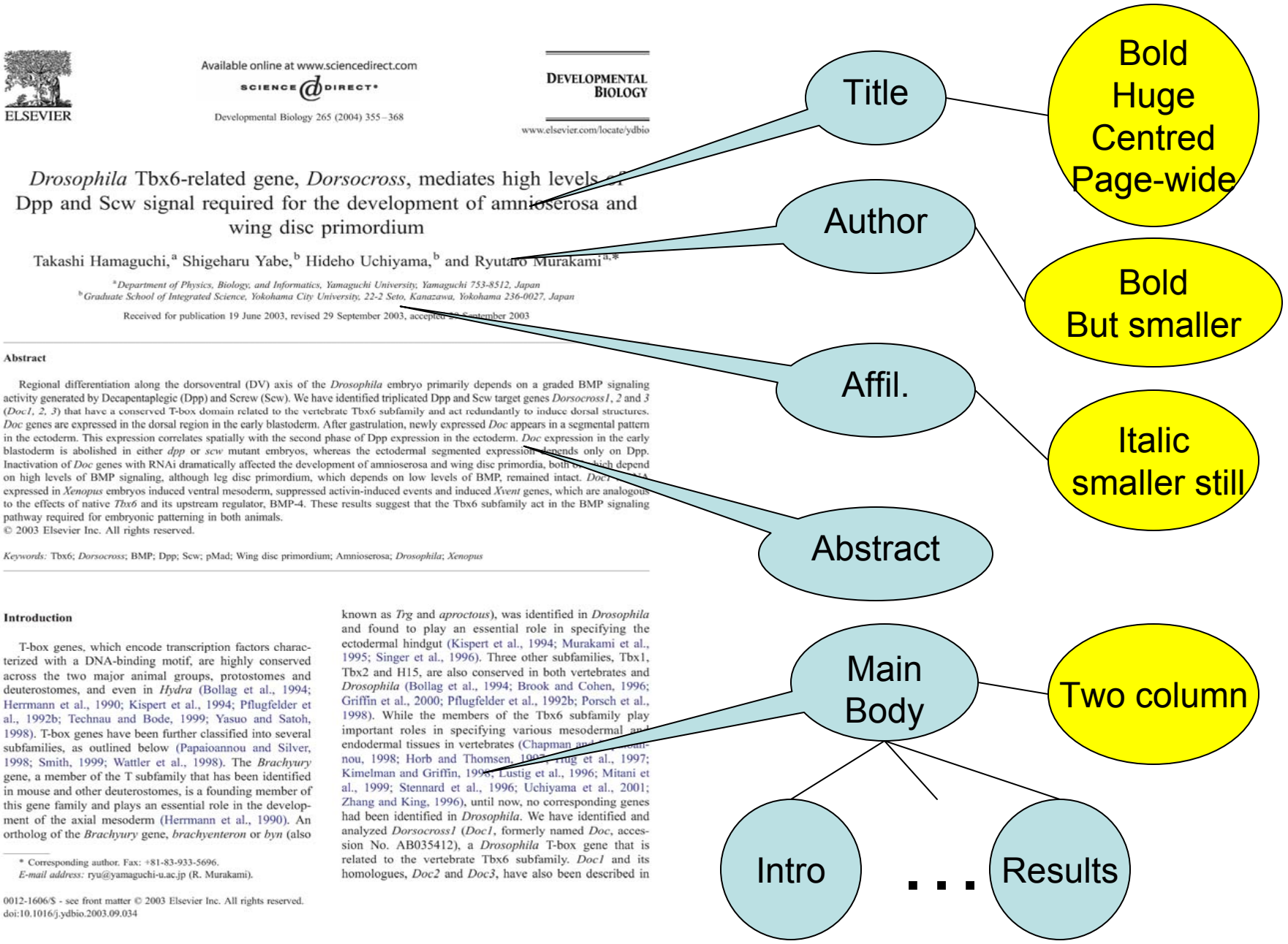
University of Cambridge Computer Laboratory
Email: ian.lewin@cl.cam.ac.uk

E-Science Context

- FlySlip: *Integrating Literature, Experiments and Curation in Drosophila Genomics Research*
  - www.cl.cam.ac.uk/users/av308/Project_Index
  - BBSRC grant no: 16291
- CitRaz: *Rhetorical Citation Maps and Domain Independent Argumentative Zoning*
  - EPSRC grant no: GR/S27832/01

Overview

- Hierarchical text structure versus Presentational structure
- Value of hierarchical text structure to NLP
- PTX: PDF to XML processing framework
- An evaluation
- Where next?



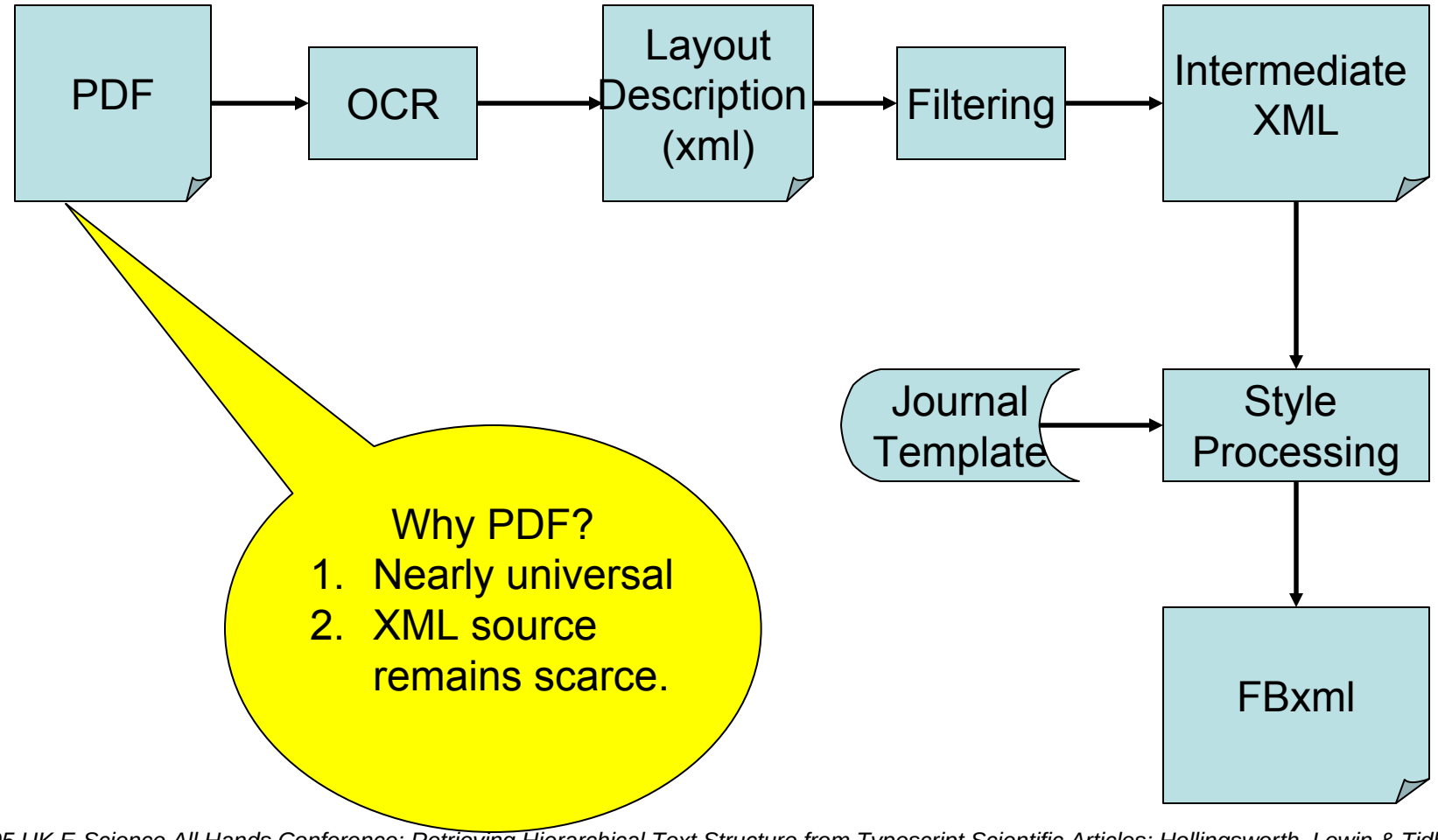
2005 UK E-Science All Hands Conference: Retrieving Hierarchical Text Structure from Typescript Scientific Articles: Hollingsworth, Lewin & Tidhar

Hierarchical Structure for NLP

- Information Retrieval, Extraction, Classification
  - Build patterns over parsed structures over words
  - Build chains of co-reference between named entities
  - Discern which structures (abstract, results, methods) patterns are matched in
  - Label parts of structure (e.g. in *Discussion,* this is Background information, this is others' results, this is new results

- Information re-presentation. Present an article back to the user but with “added value”
  - Gene names highlighted
  - Navigation by author’s sectioning
  - Navigation by repeated references
- Linkage of article to FlyBase database contents

PTX : processing framework



Why OCR?

- Standard pdf “text extraction” tends to mishandle or omit important information
- Textual characters that are encoded graphically
- Styles (bold, italics,positional info) that are important for recovering *structure.*
  - *Single line paragraphs in italics are (sub) headings*
  - *Centred italics after authors are affiliations and not the beginnings of abstracts*
  - *Small font, small size paragraphs at foots of pages not following a full stop (or other sentence delimiter) are likely footnotes*
- Styles important for content. In genomics, italicized words often indicate genes, superscript suffixes indicate alleles

Layout XML

- Documents
- Pages
- Regions
- Paragraphs
- Lines
- Words
- Characters

- Stylistic annotations on these elements
  - Font size
  - Font type
  - Indentation
  - Alignment
  - coordinates

…
<paragraph para-type="text" align="justified" left-indent="0" right-indent="72" >
<ln baseline="2511" ff="Times New Roman" fs="1000">
<wd l="1243" t="2381" r="1579" b="2520" char-attr="italic">
<ch l="1243" t="2381" r="1397" b="2515">**D**</ch>
<ch l="1402" t="2424" r="1493" b="2520">**o**</ch>
<ch l="1498" t="2424" r="1579" b="2520">**c**</ch>
</wd>
<wd l="1661" t="2419" r="2098" b="2558">
<ch l="1661" t="2419" r="1752" b="2558">**g**</ch>
<ch l="1757" t="2419" r="1838" b="2520">**e**</ch>
<ch l="1843" t="2419" r="1944" b="2515">**n**</ch>
<ch l="1944" t="2419" r="2026" b="2520">**e**</ch>
<ch l="2035" t="2419" r="2098" b="2520">**s**</ch>
</wd>
<wd l="2179" t="2419" r="2414" b="2520">
<ch l="2179" t="2419" r="2266" b="2520">**a**</ch>
<ch l="2261" t="2419" r="2333" b="2515">**r**</ch>
<ch l="2333" t="2419" r="2414" b="2520">**e**</ch>
</wd>
…

Journal Template Processing

- Generic processing: Top down, left to right traversal of the layout description
- Journal specific: attach program code that executes when certain tags (new zone, zone finishes, new page, new paragraph…) are encountered

Current….

- Programming framework that enables structure discovery by the writing of good rules
  - "a change of zone with font-style change when processing the abstract indicates that the abstract has finished"
  - "a change of zone where top-of-zone is top of page and 'high enough' and zone encompasses a one line paragraph indicates this is a page header"

Near future

- We’re moving to a more structured approach with a generic *scientific document* model (expressed as a finite state machine) and a predefined set of document features whose detection permits state transition
  - Simpler to generate new journal templates
  - Possible to learn templates from examples
  - Possible to learn common mistakes and correct

Evaluation

- Acid test will be: do the FlyBase curators find the outputs *usable*?
  - There will be mistakes; some could be crippling; others could be irrelevant
- Meantime, we can measure some coarse statistics
  - F-Measure on structure tags
  - Character recognition & style accuracy
    - Greek characters
    - Superscripts/subscripts

Evaluation

<TITLE>Proteoglycan UDP-Galactose:/3-Xylose
/31,4-Galactosyltransferase I Is Essential for
Viability in Drosophila melanogaster*
</TITLE>

<ABSTRACT>Heparan and chondroitin sulfates play
essential roles in growth factor signaling during
development and share a common linkage
tetrasaccharide structure,
GlcAβ1,3Galβ1,3Galβ1,4Xylβ1-O-Ser. In the
present study, we identified the
proteoglycan UDP-galactose:β-xylose
β1,4-galactosyltransferase I (dβ4GalTI), and
determined its substrate specificity. The enzyme
transferred a Gal to the -β-xylose (Xyl) residue,
confirming it to be the Drosophila ortholog
of human proteoglycan UDP-galactose:β-xylose
β1,4-galactosyltransferase I.

Does “abstract” occur in the same context window (upto 4 words either side) in a Gold Standard annotation?

Does β occur in the same context window in an HTML version of this published paper?

Evaluation results

- Structure:
  - Journal A (dev. set of 8 papers)
    - 18 papers from same year
      - Precision = 97.5 Recall = 96.2
    - 10 papers from previous year
      - Precision = 93.0 Recall = 94.2
  - Journal B
      - Precision = 73.9 Recall = 90.4
- Styles
  - Superscript: Precision = 100% Recall = 97.6%
  - Greek characters: Precision = 73.1%

Evaluation

- Structure Recognition – generally good
  - Many mistakes in tabular data which OCR misrecognizes as tiny textual paragraphs (looks correctable)
  - Some mistakes where citation strings in *another* English font is mis-recognized as Greek font (may be less easy to correct!)
- Superscript/subscript recognition excellent
- Greek character recognition generally good (bar citation string problem)

Future Work

- Develop generic “scientific document” model; ideally train journal specific variations via machine learning
- Find out what really matters to FlyBase database curators
- Incorporate feedback into curation cycle for iterative NLP improvement
- Incorporate images and tables