A General Model of Chinese News Content Markup

Ying-chun Hsieh Shyue-shuo Huang
Christian Wittern John Lehman
Rick Jelliffe Ching-chun Hsieh

Hong Kong
May 24 , 2004

News

Elements (5W1H)

who--subject *(person, organization)* of news event

what--happened, happening

where--place

when--time thing(s) happened

why--cause

how--situation & process

E.L.Shuman, 1894

(from Frank Luther Mott, News in America, Harvard University Press, Mass. 1952, p.158)

News Event

The occurrence which gives rise to media coverage will have fulfilled one or more, or an amalgam of ***NEWS VALUES***.

– key event

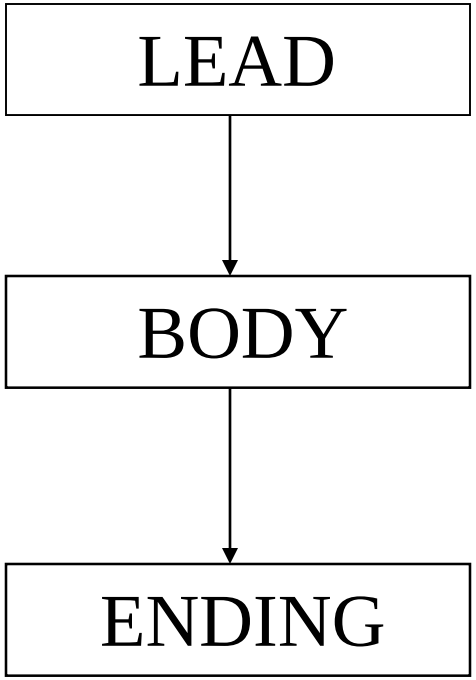
– similar event

– thematically related event

A Dictionary of Communication and Media Studies

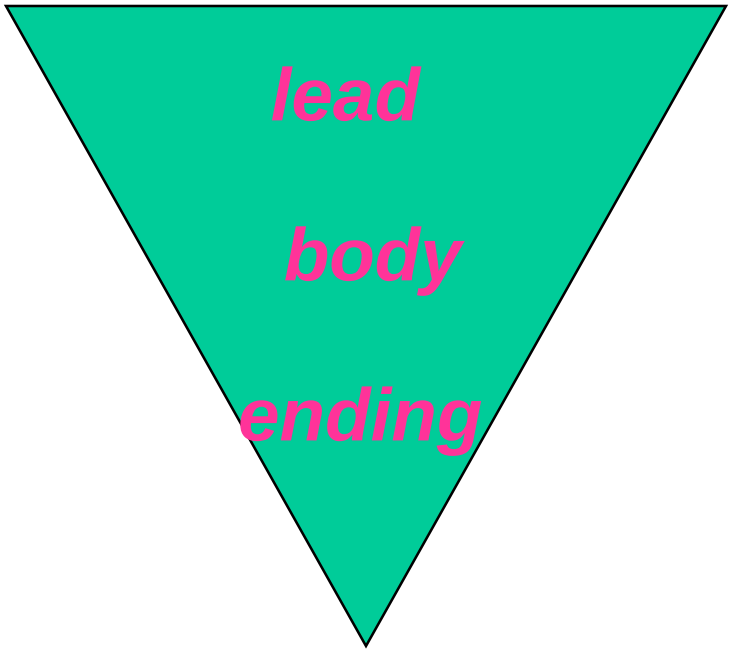
Arnold (Fourth edition), 1997, pp.78-79

News Writing Structure

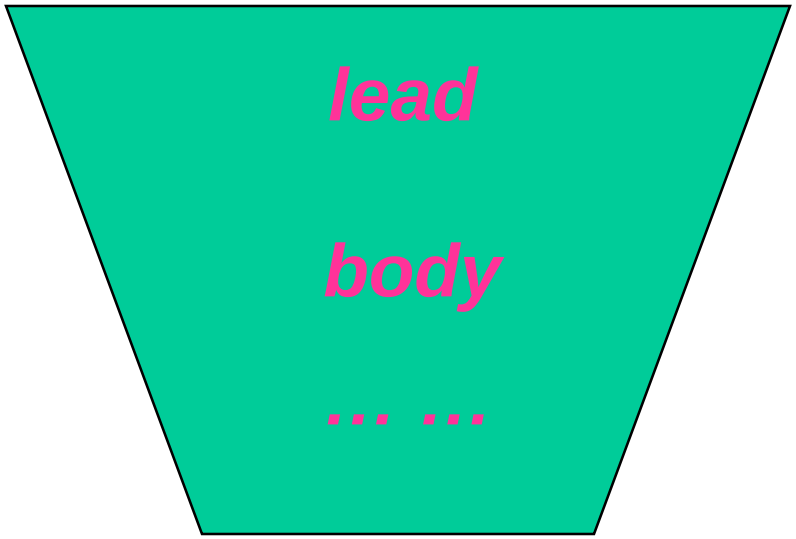


Straight News Writing

(Inverted Pyramid)



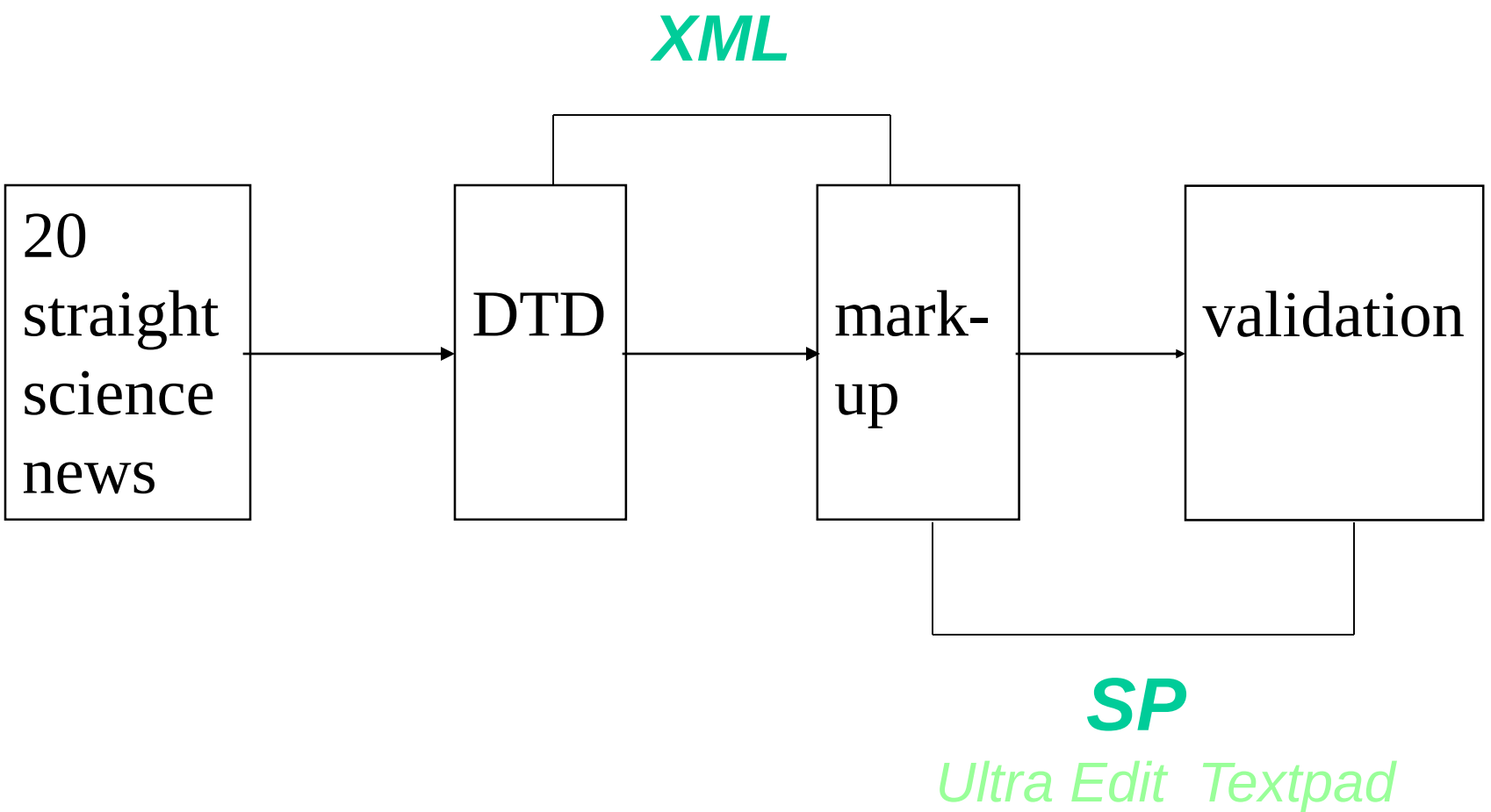
reporter



no ending

editor

Working Procedure

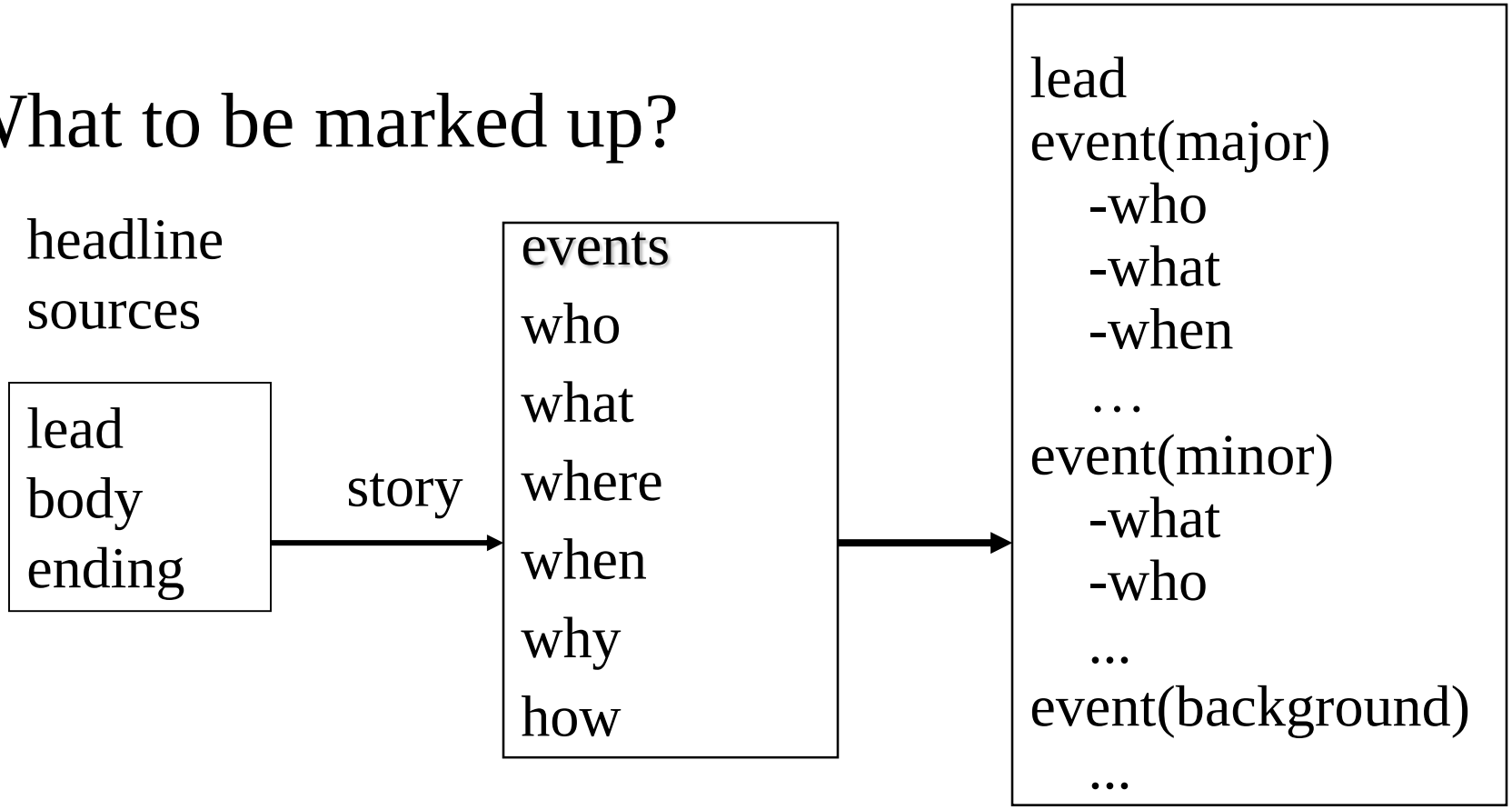


Outline

- To analyze news content (Chinese newspaper)
  - science news content (straight news)
- To mark up news content (Chinese newspaper) by XML
  - science news content, only straight news
  - no feature writing
  - no depth / investigative / special reporting
- To Mark up by semantic unit
  - instead of syntactic unit in a news context semantic unit as the basic unit

Outline

What to be marked up?



DTD *(English) I*

```
<?xml version="1.0" encoding="big5"?>
<!DOCTYPE Science_News [
<!ELEMENT Science_News (#PCDATA | Science_Story)*  >
<!--              Writing Structure                    -->
<!ELEMENT Science_Story ( headline | source | lead | body | ending)*>
<!ELEMENT headline (#PCDATA)>
<!ELEMENT source (journalist_name | newspaper_name |
                        news_agency_name | wired_service_date
                        | publishing_date | place | page)*>
<!ELEMENT journalist_name          (#PCDATA)>
<!ELEMENT newspaper_name        (#PCDATA)>
<!ELEMENT news_agency_name    (#PCDATA)>
<!ELEMENT wired_service_date      (#PCDATA)>
<!ELEMENT publishing_date          (#PCDATA)>
<!ELEMENT place                          (#PCDATA)>
```

DTD *(English) II*

```
<!ELEMENT page      (#PCDATA)>
<!ELEMENT lead      (#PCDATA | event)*>
<!ELEMENT event     (#PCDATA | who | what | when
                      | where | how | why )*>
<!ELEMENT who       (#PCDATA)>
<!ELEMENT what      (#PCDATA)>
<!ELEMENT when      (#PCDATA)>
<!ELEMENT where     (#PCDATA)>
<!ELEMENT how       (#PCDATA)>
<!ELEMENT why       (#PCDATA)>
<!ELEMENT body      (#PCDATA | event)*>
<!ELEMENT ending    (#PCDATA |event | who | what | when
                      | where | how | why )*>
```

DTD *(English) III*

```
<!ATTLIST Science_News      id          ID     #REQUIRED>
<!ATTLIST event             id          ID     #REQUIRED
                    classification  (major | minor | other |
                                    background | detailed)
                                    "detailed"
                    statement_type  (fact | opinion | mixed)
                                     "fact"
                    content_characteristic  (news_information |
                                     scientific_information)
                                    "news_information”
                    relative_events  CDATA #IMPLIED
                    relation_types   CDATA #IMPLIED      >
```

DTD *(English) IV*

```
<!ATTLIST who          id      ID #REQUIRED>
<!ATTLIST what         id      ID #REQUIRED>
<!ATTLIST when         id      ID #REQUIRED>
<!ATTLIST where        id      ID #REQUIRED>
<!ATTLIST how          id      ID #REQUIRED>
<!ATTLIST why          id      ID #REQUIRED>
<!--End of NEWS DTD-->
]>
```

What is the purpose ?

- Newspaper retrieval (long-term)
  - readers
  - researchers
  - teachers
  - others

What is the purpose ? II

- Information exchange (long-term)
  - between / among organizations, individuals, computers
  - communication of different languages

What is the purpose ? III

- Newspapers / news printing media
- Newsmen(women) / journalists / editors
  - checking writings
- Teachers / professors / students
  - teaching, learning
- Researchers
  - content analysis
- Experience sharing with XML / TEI groups

Limitation

- Shortage of Chinese XML / TEI experts
- Lack of semantic analysis of news content in Chinese (only few references)
- No financial support

Desiderata

- more discussion / help from XML / TEI experts
- more information sharing with news groups / journalism academic community; not necessary in Chinese / Taiwan only, but also in other languages / countries.
- with some financial aids for hiring research assistants and other needs.

Next stage *(phase)* I

- test the system with the database of the NDAP Newsletter in the future.
- apply to other kinds of news content, such as features, special reporting, columns, commentaries, etc.

Next Stage *(phase)* II

- extend to political / economic / criminal news, etc.
- develop a mature recursive model to represent the relationship among events, who, what, where, when, why and how.

The End

Thank you for your attention!