questions for segmenting:

Here are some "principles" in segmenting proposed in some books
(for the sake of simplicity, assume a string has two morphemes):

treat the string as a word if one or more of the following hold:
- the meaning of the two morphemes is noncompositional,

- one or both morphemes can not stand alone

- one morpheme's meaning is altered or de-emphasized

- the string is so commonly used that it has become a fixed combination.

Our questions are:
are the "principles" correct? and are they adequate for the segmenting task?

The following is a list of phrases from the 100
sentences that might be problematic for segmenting:

1. affix: what tests should be used to decide whether a string of characters is an affix or not?

   most systems treat an affix as part of a word.

   examples of affixes:
   化 in 一体化
   反 in 反 倾销
   制 in 结汇 制

   are the following affixes?
   省 in 河北 省,
   市 in 呼和浩特 市/呼 市(呼 is the short form of 呼和浩特),
   县 in 通 县(a county outside Beijing)
   (does the number of syllables of the word before 县 etc. matter?)

   点 in 三 点, 月 in 三 月, 年 in 1998 年, 号 in 30 号

   主义 in 实用 主义, 制度 in 公务员 制度

   副 in 副 行长
   局 in 教育 局, 厂 in 钢铁 厂
   人 in 残疾 人
   社 in 新华 社
   率 in 年 均 增长率

军训/部　使命/感 紧迫/感　大奖/赛

2. numbers

most systems treat numbers such as 1234 as a word

but for larger units, should they be segmented?
- affixes or words: e.g.
  几/来/多 in 30 几 万
  上下/左右 in 30 万 上下
  约/大约 in 约 30 万
  第 in 第 三

- a sequence of number characters: one word or not?
  30 万
  三十万
  三 分之 一

- punctuation: one word or not?
  3 . 8
  45 %
  ６ ： ２

3. date and time

one word or not?
点 in 三 点, 月 in 三 月, 年 in １９９５/年, 号 in 30 号
日 in ２３ 日

4. Nouns
(1) names
- should Chinese personal names be segmented?
  e.g. 王荣生

- should foreign personal names be segmented?
  e.g. 圣 . 玛丽
  鲁珀特・比林纳姆

- organization names: should it be segmented?
  e.g. 中国船舶工业总公司(中国/船舶/工业/总公司)
  北大 方正
  中国人民银行(中国/人民/银行)
  中国/航空/器材/公司
  中央银行
  全国/人大
  西南/得克萨斯大学(西南/得克萨斯/大学)
  上海/第一医科大学(上海/第一/医科/大学)
  长江/轮船/海外/公司
  旭峰旅业发展有限公司
  北方/电讯/公司
  ＤＭＳ/程控/交换机/制造厂

贝尔/北方/研究所
联合国/安理会

- acronym:
  e.g. 新澳
  东新澳
  中/美 中/泰
  亚/欧

- title
  乔/委员长 贾海燕/博士

(2) others

- modifier+noun, should they be segmented?
  e.g. 高山, 求医 船 ９０/年代 亚太/地区 东南亚/新澳
  科技/人员 年均/增长率(年/均/增长/率)
  工作/量 大/项 元/月 发展/史
  少数/民族 年/产量 成套/设备(成/套/设备)
  大/奖赛 早餐/会
  劳动/产品 进出口/贸易(进/出口/贸易) 中/方 高/层
  摄影/艺术 花样/游泳 规定/动作 自选/动作
  学术/交流 常任/理事国 领导/人 剑桥/大学
  积极/作用 记者/招待会 誓师/会
  运动/队 共同/市场 股票/交易所 证券/交易/所 长江/三峡
  促销/团 政治/体制 大/奖/赛 中华/民族
  有限/公司 暴力/行动 世界/文化 民族/文化 北大/方正
  民族/文化 随军/妓女 经济/援助
  三角/战 清晰/度 太子/港 世界/银行 船/民 国际/舞台
  国际/有限公司 高/清晰度 多/国/部队 三峡/风光
  朝鲜/劳动党 四/国 给惠/国家
  新/中国 除草/剂 发达/国家 最惠国/待遇 财务/信托/公司
  国家/计委 谅解/备忘录 超大/规模/集成电路
  北京/邮电/学院 实验/室 律师/事务所 新/路
  最后/通牒 德/军 关贸/总协定 宏观/调控
  大/洲 大/洋

5. verb and verb compounds: should verb compounds be treated as words?
- reduplicate:
  e.g. 喜欢 不 喜欢
  喜 不 喜欢
  看不见
  看得见

- verb+object compounds where other components can be inserted between the verb and the object

  e.g. 吃饭 睡觉
  造/船
  发/稿
  获/奖
  进/港 避/风 交/权 组/队

驻/华

- resultative verb compounds:
  e.g. 研制 成功
  踢 破, 打 破
  洗 干净

- directional verb compounds:
  e.g. 走 进来

- other compounds
  e.g.
  坐 在
  流/入
  迈/向 指/出 有利/于
  送/到 进/入 处/于 发/出
  感觉 到 感/到 来/自
  驶 至
  团结 在
  留在
  做出
  联结 起来
  低/于
  找/到

  敬/献

6. adverbs:

   e.g. 从 不
   极 大 in 极 大 地 促进
   较/低
   激动/地 很/快 不/久

7. short form:
   远南, 随军, 预警
   年 均 增长率

8. idioms, words from classic Chinese, should they be segmented?
   真相 大白, 各 具 特色 华夏/之/光
   在内, 为首, 取向, 其中, 有所, 给惠
   事 隔 in 事 隔 两 年
   训练/有素(训练/有/素)
   秩序/井然
   一/比/高低

9. strings which have different meanings when treated as phrases and as words. How to decide whether they are words or not in a context?

   有 可能, (v+n) as a phrase, adv as a word
   很多, (adv+adj) as a phrase, det as a word

只能 不能 (adv+aux) as a phrase, CC as a word,
没有: (adv+v) as a phrase, adv as a word
不久: (adv+adj) as a phrase, adv as a word
不/够

10. phrases that are so commonly used that they have become a fixed combination: should they be segmented?
及其

根本上 in 从 根本上
各国 本国
各个
这些
而又 in 热烈 而又 庄严
不到 in 不到 五 分钟
不足 in 不足 ０・５ 美元
特别 是 in 特别 是 在 ... 中 所 起 的 积极作用
从小到大
我/方
全/岛 一/些
这/些
越/来/越(越来/越重要)
之/间 本/月 本/次 更/多 近/年/来 进/一/步 这/次
最/新/消息 海内/外 不/同 为/了 先/后
仅/有 迄/今