

HƯỚNG DẪN NHẬN DIỆN ĐƠN VỊ TỪ TRONG VĂN BẢN TIẾNG VIỆT

Nguyễn Thị Minh Huyền, Hoàng Thị Tuyên Linh, Vũ Xuân Lương

Báo cáo SP8.2

I. Nguyên tắc tách từ

1. Hướng tới chuẩn tách từ - ISO/TC37/SC4/WG2/WordSeg

Trong các hoạt động về chuẩn hoá tài nguyên ngôn ngữ của ISO/TC37/SC4 có nhóm làm việc WG2/WordSeg[1-3] về vấn đề chuẩn hoá tách từ cho các ngôn ngữ trong đó ranh giới giữa các từ không thể xác định rõ ràng chỉ dựa vào hình thức in ấn (như sử dụng dấu cách trong tiếng Anh).

Cho đến nay, nhóm làm việc này đã đưa ra một số bản thảo (trang web <http://tc37sc4.org>) hướng dẫn nguyên tắc chung về việc đưa ra chuẩn tách từ.

2. Đặc trưng cấu tạo từ tiếng Việt

Các phương thức cấu tạo từ tiếng Việt:

Từ đơn:

Từ có ý nghĩa từ vựng.

Từ có ý nghĩa ngữ pháp (từ công cụ).

Từ tượng thanh.

Từ cảm thán.

Từ phức:

Từ ghép.

- Từ ghép đẳng lập (tổng hợp).

- Từ ghép chính phụ.

- Từ ghép phụ gia (yếu tố ghép trước hay ghép sau để tạo từ hàng loạt).

Từ láy.

Dạng lặp.

Ngữ cố định:

Thành ngữ (cao chạy xa bay, tránh vỏ dưa gặp vỏ dừa...).

Quán ngữ (nói tóm lại, đáng chú ý là, mặt khác thì...).

Ngoài ra, trong văn bản còn có các thành phần sau:

Tên riêng (người, địa danh, tổ chức).

Các dạng ngày – tháng – năm.

Các dạng số – chữ số – kí hiệu.

Dấu câu, dấu ngoặc.

Từ tiếng nước ngoài.

Chữ viết tắt.

3. Đề xuất nguyên tắc tách từ cho tiếng Việt

Nguyên tắc tách từ cho tiếng Việt xét các loại đơn vị từ vựng sau đây:

Từ đơn.

Từ ghép đẳng lập.

Từ ghép chính phụ.

Từ ghép phụ gia (kết hợp với yếu tố cấu tạo từ: *bất, vô, hoá, phi, viên*, v.v.).

Từ láy, dạng lặp.

Thành ngữ.

Quán ngữ.

Tên riêng.

Ngày – tháng – năm, số – chữ số – kí hiệu.

Dấu câu, ngoặc.

Từ tiếng nước ngoài.

Chữ viết tắt.

II. Hướng dẫn cụ thể

Coi là một đơn vị từ khi thực hiện tách từ đối với các đơn vị có những đặc điểm sau đây:

1. Từ đơn.

a. Từ đơn là thực từ:

- Những từ một tiếng có ý nghĩa từ vựng độc lập, có chức năng định danh (gọi tên các sự vật, hiện tượng, hành động, phẩm chất, thuộc tính, quan hệ trong thực tại khách quan).

- Đa số đều nằm trong vốn từ cơ bản của tiếng Việt, đã có từ lâu đời: *cha, mẹ, chân, tay, cơm, nước, lợn, gà, ăn, uống, cười, nói, xấu, đẹp*, v.v.; hoặc những từ gốc Hán hay gốc Ấn-Âu đã được Việt hoá: *tim, gan, buồng, phòng, cón, xăng, xăm, lớp*, v.v.; hoặc những từ Hán-Việt được dùng độc lập (do không có từ thuần Việt đồng nghĩa tương đương): *tuyệt, bút, học, đáp, cao, thấp*.

- Có một số vốn là dạng nói tắt của từ ghép: *rô* (cá rô), *chim* (cá chim), *thu* (cá thu), *nhụ* (cá nhụ), *đé* (cá đé), v.v.

b. Từ đơn là hư từ:

- Những từ một tiếng không có ý nghĩa từ vựng độc lập, không có chức năng định danh.
- Không có khả năng độc lập làm thành phần câu.
- Dùng để biểu thị các quan hệ ngữ pháp giữa các thực từ.
- Gồm phụ từ, liên từ, giới từ: *đã, sẽ, đang, vừa, mới, từng, vẫn, là, của, bằng, vì, bởi, cùng, với, nếu, tuy, nên, v.v.*

c. Từ đơn là từ tình thái:

- Những từ một tiếng đã mất ý nghĩa từ vựng và ý nghĩa ngữ pháp cụ thể, có chức năng như một phương tiện biểu thị tình thái.
- Không có khả năng độc lập làm thành phần câu.
- Biểu thị mối quan hệ giữa người nói với thực tại phát ngôn.
- Gồm thán từ và trợ từ: *à, ư, nhi, nhé, ời, hử, sao, a, ơ, ối, ái, thế, nào, đâu, vậy, v.v.*

2. Từ ghép đẳng lập

- Do hai thành tố (A và B) có ý nghĩa thực kết hợp với nhau theo quan hệ bình đẳng về nghĩa.
- Hai thành tố bao giờ cũng thuộc cùng một phạm trù ngữ nghĩa hoặc có quan hệ logic với nhau.
- Trật tự giữa hai thành tố nói chung có thể thay đổi được (AB hoặc BA): *quần áo – áo quần, chung riêng – riêng chung, đỏ đen – đen đỏ, ốm đau – đau ốm, v.v.*

2.1. Từ ghép đẳng lập gốc Việt

- Từ ghép đẳng lập gốc Việt là từ ghép trong đó hai thành tố đều là từ gốc Việt.

a. Từ ghép đẳng lập gốc Việt gồm hai thành tố có sự *gần nhau về nghĩa*:

đất nước – trời đất – đất cát – ruộng đất – ruộng vườn – ruộng nương; ẩm chén, bát đĩa, bó con, cày cuốc, chồng con, cướp phá, dệt thêu, làng xã, lúa gạo, nương vườn, râu tóc, tài sức, thác ghềnh, thầy cô, thiếu kém, thu đông, vá may, vải sợi, vườn trại, xinh đẹp, v.v.

b. Từ ghép đẳng lập gồm hai thành tố có sự *trái nhau về nghĩa*:

đỏ đen, may rủi, trong ngoài, trước sau, trên dưới, thảo lẫn, cao lớn, chung riêng, hay dở, khen chê, v.v.

2.2. Từ ghép đẳng lập gốc Hán

- Từ ghép đẳng lập gốc Hán là từ ghép trong đó hai thành tố đều là từ gốc Hán.

a. Từ ghép đẳng lập gốc Hán gồm hai thành tố đã được Việt hoá hoàn toàn (được dùng độc lập như những từ gốc Việt khác):

ân nghĩa, công tư, đầu não, đấu tranh, học tập, lợi lộc, thuận lợi, v.v.

b. Từ ghép đẳng lập gốc Hán gồm hai thành tố chưa được Việt hoá hoàn toàn (không dùng độc lập như những từ gốc Việt khác):

chung thủy, giang sơn, kiến thiết, mỹ lệ, quốc gia, tao nhã, tranh chấp, v.v.

c. Ngoài ra còn có những từ ghép đẳng lập gồm một thành tố gốc Việt và một thành tố gốc Hán (in nghiêng là gốc Hán):

binh lính, bụng *dạ*, gan *dạ*, lính *tráng*, nuôi *dưỡng*, v.v.

3. Từ ghép chính phụ

- Do hai thành tố (A và B) trực tiếp kết hợp với nhau theo quan hệ không bình đẳng. Đó là sự phối hợp giữa một thành tố chính có ý nghĩa khái quát (A) và một thành tố phụ (B) có ý nghĩa hạn định.

- Ý nghĩa từ vựng do thành tố chính (A) quyết định; thành tố phụ (B) có vai trò bổ sung, phân loại, chuyên biệt hoá, sắc thái hoá cho thành tố chính.

- Thành tố A có thể dùng thành từ, còn thành tố B thì có thể không có tư cách ngữ pháp đó. Trật tự giữa hai thành tố A và B là không thể thay đổi được. So sánh: *xe máy* – *máy xe*; *không quân* – *quân không*, v.v.

3.1. Từ ghép chính phụ gốc Việt

- Vị trí của hai thành tố A và B trong cấu tạo từ ghép chính phụ gốc Việt là *chính trước* – *phụ sau* (AB: *xe máy*, *xe đạp*, *xe tăng*).

a. Từ ghép chính phụ bậc 1, trong đó thành tố A là từ đơn và thành tố B là một từ đơn, hoặc một từ ghép, hoặc một tổ hợp từ:

+ *cá* (A): *cá mè*, *cá rô*, *cá trắm*, *cá quả*, *cá hồng*, *cá voi*, *cá heo*, *cá chai*, *cá bột*, *cá nhà táng*, *cá săn sắt*, *cá thồn bơn*, v.v.

+ *chim* (A): *chim gáy*, *chim khuyên*, *chim ngói*, *chim hát bội*, *chim cánh cụt*, *chim phượng chèo*, *chim thầy bói*, v.v.

+ *hoa* (A): *hoa hồng*, *hoa nhài*, *hoa lan*, *hoa li*, *hoa sói*, *hoa mõm sói*, *hoa mép dê*, *hoa cút lợn*, *hoa loa kèn*, v.v.

+ *rau* (A): *rau má*, *rau sam*, *rau răm*, *rau sống*, *rau húng*, *rau thơm*, *rau tập tàng*, v.v.

+ *cà* (A): *cà chua*, *cà bát*, *cà pháo*, *cà tím*, *cà dái dê*, *cà độc dược*, v.v.

+ *máy* (A): *máy bay*, *máy bơm*, *máy sát*, *máy xay*, *máy kéo*, *máy cày*, *máy gặt đập*, *máy phát điện*, *máy quay đĩa*, *máy thu hình*, v.v.

+ *xe* (A): *xe đạp*, *xe tăng*, *xe cút kít*, *xe cứu hoả*, *xe cứu hộ*, *xe cứu thương*, v.v.

+ *bếp* (A): *bếp dầu*, *bếp điện*, *bếp gas*, *bếp từ*, v.v.

+ *nồi* (A): *nồi hầm*, *nồi hấp*, *nồi hơi*, *nồi supde*, *nồi áp suất*, *nồi cơm điện*, v.v.

+ *bàn* (A): *bàn đọc*, *bàn giấy*, *bàn thờ*, *bàn cờ*, v.v.

+ *làm* (A): *làm bếp*, *làm biếng*, *làm công*, *làm giàu*, *làm việc*, v.v.

+ *đen* (A): *đen đúa*, *đen giòn*, *đen hắc*, *đen ngòm*, *đen nhem*, *đen sì*, v.v.

v.v....

b. Từ ghép chính phụ bậc 2, trong đó thành tố A là một từ ghép và thành tố B là một từ đơn, hoặc một từ ghép (gốc Việt hoặc gốc Hán), hoặc một tổ hợp từ:

+ cá mè (A): cá mè hoa, cá mè trắng, v.v.

+ máy bay (A): máy bay bà già, máy bay trực thăng, máy bay lên thẳng, máy bay cường kích, máy bay khu trục, máy bay không người lái, v.v.

+ máy xay (A): máy xay sinh tố, máy xay thịt, v.v. (???)

+ động cơ (A): động cơ diesel, động cơ đốt trong, động cơ điện, động cơ vĩnh cửu, v.v.

v.v...

3.2. Từ ghép chính phụ gốc Hán

a. Trường hợp thông thường, hai thành tố A và B trong từ ghép chính phụ gốc Hán được sắp đặt theo trật tự *phụ trước – chính sau*. Trong đó, thành tố A là từ đơn được dùng độc lập hoặc không độc lập và thành tố B là một từ đơn, hoặc một từ ghép.

+ ca (A): dân ca, đồng ca, xướng ca, khai hoàn ca, v.v.

+ dân (A): bình dân, cư dân, ngư dân, nông dân, v.v.

+ học (A): bác học, văn học, kinh tế học, cổ sinh vật học, v.v.

- Chú ý: Có trường hợp thành tố B là từ gốc Việt, gốc Anh.

môi hoá, nhót ké, ampe ké, logic học, v.v. (*môi, nhót, ampe, logic* là B)

b. Có trường hợp hai thành tố A và B trong từ ghép chính phụ gốc Hán được sắp đặt theo trật tự *chính trước – phụ sau*; trường hợp này A là động từ và B là từ đơn gốc Hán được dùng độc lập hoặc không độc lập.

+ đá (A): đá đảo, đá động, đá kích, đá phá, v.v.

+ thuyết (A): thuyết giảng, thuyết lí, thuyết minh, thuyết phục, v.v.

CHÚ Ý:

1. Với loại từ ghép chính phụ, khi thành tố A là danh từ chỉ đồ vật (vật vô sinh: *máy, xe, bếp, nồi, bàn*, v.v.), thì thường thành tố B là động từ hoặc tính từ biểu thị ý nghĩa công dụng, mục đích, cách thức, tính chất. Theo đó, các tổ hợp kiểu: *nồi đồng* (*nồi* bằng *đồng*), *nồi đất* (*nồi* bằng *đất*), *mâm nhôm* (*mâm* bằng *nhôm*), *bàn gỗ* (*bàn* bằng *gỗ*), *ghế đá* (*ghế* bằng *đá*), v.v. không có tư cách là một từ ghép chính phụ. B ở đây (*đồng, đất, nhôm, gỗ, đá*) là những danh từ chỉ chất liệu.
2. Trong tiếng Việt còn có những từ có nhiều tiếng (bao gồm cả từ vay mượn đã được Việt hoá, hoặc có hình thức phiên âm gần giống với tiếng Việt), xét theo phương thức cấu tạo thì không thuộc loại từ ghép cũng không thuộc loại từ láy. Chúng bao gồm những tiếng không có nghĩa hoặc mờ nghĩa (có thể do chưa biết được nghĩa gốc), phải cả khối gồm nhiều tiếng hoà quyện làm một chỉnh thể chặt chẽ mới có nghĩa: *bỏ nông, bỏ hóng, bù nhìn, mặt chược, ca la thầu, ba lô, béc giê, cà phê, căng tin, xi măng, xích lô*, v.v. Những từ này cũng được xếp chung vào nhóm từ ghép.
3. Cũng coi là từ ghép với các tổ hợp gộp (của hai từ ghép) biểu thị ý nghĩa tổng hợp:
 - Kết hợp giữa hai, ba thành tố đầu trong mỗi từ ghép: công nông (*công nhân* và *nông dân*), công nông binh (*công nhân, nông dân* và *binh lính*), v.v.

- Cả hai từ ghép đều có chung thành tố chính A (đứng cuối): *y bác sĩ* (*y sĩ* và *bác sĩ*), *ưu nhược điểm* (*ưu điểm* và *nhược điểm*), *khám chữa bệnh* (*khám bệnh* và *chữa bệnh*), *binh công xương* (*binh xương* và *công xương*), v.v.

- Dạng viết đầy đủ: *phòng cháy chữa cháy*, *phòng bệnh chữa bệnh*, v.v.

4. Trong những trường hợp lưỡng lự có thể xét đến các lí do sau đây:

a) Những tổ hợp có cấu tạo tương đương như các từ đã được thu thập trong *Từ điển công cụ* (từ điển dùng làm công cụ tách từ), nhưng không được hoặc chưa được thu thập (trong ngoặc là từ có trong *Từ điển công cụ*):

anh hồn (*anh linh*), *chao ơi* (*chao ôi*), *chúng bay* (*chúng mày*), *chúng nó* (*chúng tôi*, *chúng ta*), *con ở* (*người ở*), *công dân quyền* (*quyền công dân*), *đành tâm* (*đang tâm*), *đôi lúc* (*đôi khi*), *giời ơi* (*trời ơi*), *giời phật* (*trời phật*), *hai thân* (*song thân*), *khăn tay* (*khăn mùi soa*), *khôn nổi* (*khốn một nổi*), *không thể nào* (*không thể*), *luật phép* (*luật pháp*), *oai tín* (*uy tín*), *quan binh* (cũ, như *quan quân*), *sốt tiết* (*điên tiết*), *sức của* (*vật lực*), *sức người* (*nhân lực*), *tấm gương* (như *gương*), *thang thuốc* (*thuốc thang*), *tín tâm* (*lòng tin*), *thiệt ra* (*thật ra*), *tổng sản phẩm trong nước* (*tổng sản phẩm quốc nội*), *xem trọng* (= *coi trọng*), v.v.

b) Chưa được thu thập trong *Từ điển công cụ*, nhưng đã được thu thập ở một vài quyển từ điển khác:

giá trị gia tăng (NLân), *khách hàng* (TĐ2008), *khu công nghiệp* (TĐ 2008), *kiến trúc sư trưởng* (NLân), *kim tiêm* (Đại TĐ, NLân, VTân), *lưu toan* (NLân, VTân), *nghe nga* (NLân), *nhà ở* (NLân, VTân), *như vậy* (NLân, VTân, LVĐức, TNghị), *như thế* (NLân, VTân, LVĐức, KTrí, TNghị), *quan binh* (LVĐức), *quan tư* (Đại TĐ, VTân, KhTrí, ĐVTập), *quốc công tiết chế* (Đại TĐ, NLân, VTân), *rẻ rẻ* (LVĐức, ĐVTập), *thù hiềm* (LVĐức), *tự đại* (ĐVTập, TNghị, LVĐức), v.v.

c) Đơn vị từ vựng mới xuất hiện (từ mới hoàn toàn, hoặc từ cũ nay dùng lại):

ảnh điểm, *bưu báo*, *bo chủ*, *giả lập*, *lục sì*, *máy để bàn*, *máy tính để bàn*, *nguyên lão nghị viện*, *quan năm*, *tác vụ*, *trình khách*, *tư túng*, v.v.

d) Các cụm từ kiểu: *đáp lễ*, *đến nỗi*, *làm sao*, *như ai*, v.v.

4. Từ láy, dạng lặp

4.1. Từ láy

- Từ láy phổ biến là từ gồm hai tiếng (song tiết, hai âm tiết), trong đó một tiếng có hình thức lặp lại âm của tiếng kia. Các tiếng kết hợp với nhau vừa có sự hài hoà về ngữ âm, vừa có giá trị biểu cảm, gợi tả.

- Thường chỉ có một tiếng có nghĩa và một tiếng mờ nghĩa: *chậm chạp* (*chậm* có nghĩa), *long lanh* (*long* có nghĩa), *lúng túng* (*túng* có nghĩa), *long tong* (*tong* có nghĩa); hoặc cả hai tiếng đều mờ nghĩa: *khấp khểnh*, *lênh đênh*, *lênh khênh*, *lêu nghêu*, *lung linh*, v.v.

a. Kiểu AA' (A là tiếng gốc, tiếng chính; A' là tiếng láy của A):

chậm chạp, *lành lặn*, *nhánh nhàu*, *vừa vặn*, v.v.

b. Kiểu A'A (A là tiếng gốc; A' là tiếng láy của A):

b.1. *bành bạch, bì bạch, long tong, lộp bộp, lúng túng, rôm rộp*, v.v.

b.2. *đềm đẹp, đo đỏ, lành lạnh, nho nhỏ*, v.v.

c. Kiểu AA:

c.1. Lặp hoàn toàn âm của tiếng gốc, phần lớn là từ tượng thanh: *ào ào, âm âm, au au, ặc ặc, âm âm, bành bành, đập đập, êm êm* (không phải tượng thanh), *ha ha, khao khao, khặc khặc*, v.v.

c.2. Lặp hoàn toàn âm của tiếng gốc một cách đơn giản (nghĩa không biến đổi gì nhiều): *cau cau, chau chau, đen đen, lẩm lẩm, quen quen, run run, xanh xanh*, v.v.

d. Kiểu ABB (B là thành tố của từ ghép chính phụ AB):

đen sì sì, đỏ lòm lòm, nông choèn choèn, tối om om, xanh lè lè, v.v.

e. Kiểu AB'B (B' là tiếng láy của B; AB là từ ghép chính phụ):

đen trùi trùi, đỏ hoen hoét, đỏ hon hỏn, cao lêu nghêu, dài đuôn đuôn, v.v.

f. Kiểu ABC (có sự biến đổi về thanh điệu) – nghiên cứu thêm:

dừng dừng dưng, sạch sành sanh, v.v.

g. Kiểu AA'AB (A là tiếng đầu của từ ghép AB; A' là tiếng láy của A; A' có cấu tạo dạng *xa*, trong đó *x* là phụ âm đầu của A, *a* là phần vần có giá trị hoà phối ngữ âm cho cả khối):

ấm a ấm ỨC, đủng đa đủng đĩnh, long la long lanh, nhí nha nhí nhảnh, v.v.

CHÚ Ý:

1. Các kiểu b.2 (của b), c.2 (của c), d, e, f, g có tài liệu phân thành *dạng láy*. Khái niệm “dạng láy” không chỉ ra được sự khu biệt với khái niệm “láy”. Vả lại, *láy* bản thân là một *dạng* của phương thức cấu tạo từ, cũng như *ghép, lặp*. Vì những lẽ đó, tài liệu này không phân biệt *từ láy* và *dạng láy* của từ.

2. Các tổ hợp dạng *ba ba, cào cào, châu chấu, chuồn chuồn*, (quả) *đu đủ*, (quả) *su su, thần lằn, thường lỗng*, v.v. xét về mặt ý nghĩa, chúng không có giá trị biểu cảm, gọi tả như các từ láy, nhưng xét về hình thức ngữ âm thì chúng có cấu tạo giống như từ láy, vì vậy tài liệu này xếp chung vào loại từ láy.

4.2. Dạng lặp

a. Kiểu AA (lặp hoàn toàn tiếng gốc để chỉ số lượng nhiều, hoặc chỉ mức độ cao; cả hai thành tố đều là danh từ):

ai ai, đầu đầu, đêm đêm, lớp lớp, ngày ngày, người người, nhà nhà, sáng sáng, tháng tháng, tối tối, v.v.

b. Kiểu AAA (thường là tượng thanh):

ầm ầm ầm, ha ha ha.

c. Kiểu AABB (AB là từ ghép đẳng lập, trong đó A ngược nghĩa với B)

đi đi lại lại, hư hư thực thực, lên lên xuống xuống, quần quần áo áo, ra ra vào vào, v.v.

d. Kiểu ABAC (B và C thường tạo thành từ ghép đẳng lập, trong đó B ngược nghĩa với C, nhưng đôi khi cũng có thể B đồng nghĩa với C; A là yếu tố chen vào đầu và giữa tổ hợp BC).

chạy ngược chạy xuôi, chẳng nói chẳng rằng, dần đi dần lại, đá đi đá lại, đảo đi đảo lại, khát quanh khát quẩn, khoắng lấy khoắng để, khua đi khua lại, người này người nọ, trông trước trông sau, về lâu về dài, v.v.

5. Từ ghép phụ gia

- Đây là kiểu tạo từ hàng loạt bằng cách ghép các yếu tố có khả năng cấu tạo từ cao (như *bất, vô, phi...*) vào trước hay sau một từ ghép khác. Có một số tổ hợp được tạo ra từ phương thức này do không có sự ổn định cao nên có thể chưa được thu thập trong từ điển giải thích ngôn ngữ, chẳng hạn *cố bộ trưởng, cựu bộ trưởng, cố giáo sư, nguyên giáo sư, v.v.*

5.1. Danh sách các yếu tố đứng trước

bán + N = N: *bán bình nguyên, bán nguyên âm, bán sơn địa, bán thành phẩm.*

bán + A = A: *bán tự động, bán vũ trang.*

bất + A = A: *bất bình đẳng, bất đắc chí, bất hợp lí, bất khả thi.*

bất + V = V: *bất bạo động, bất hợp tác.*

bất + N = N: *bất đẳng thức, bất động sản, bất phương trình.*

cố + N = N: *cố bộ trưởng, cố giáo sư, cố thủ tướng, v.v.*

cựu + N = N: *cựu bộ trưởng, cựu giám đốc, cựu thủ tướng, v.v.*

đa + N = N: *đa phương tiện, đa tác vụ*

đại + N = N: *đại bản doanh, đại bộ phận, đại công nghiệp, đại gia đình.*

hữu + N = A: *hữu hạn, hữu hình, hữu sự, hữu thần.*

hữu + V = A: *hữu dụng, hữu khuynh, hữu sinh, hữu trách*

liên + N = N: *liên bang, liên bộ, liên ngành, liên cầu khuẩn, liên chi uỷ, v.v.*

nguyên + N = N: *nguyên bộ trưởng, nguyên thủ tướng, nguyên trưởng phòng, v.v.*

nhà + V = N : *nhà cung cấp, nhà phê bình* (chú ý: tách phần bổ ngữ tiếp sau, nếu có: *nhà phê bình / văn học ; nhà phê bình / điện ảnh, ...*).

phi + N = A: *phi chính phủ, phi lợi nhuận, phi nhân đạo, phi nông nghiệp, phi windows.*

phó + N = N: *phó chủ nhiệm, phó chủ tịch, phó viện trưởng, phó giám đốc.*

siêu + N = N: *siêu giai cấp, siêu hạng, siêu sao, siêu cầu thủ, siêu lợi nhuận, siêu trầm*

siêu + V = V: *siêu dẫn, siêu thoát, siêu thắng*

siêu + A = A: *siêu thực, siêu trường, siêu trọng*

tái + V = V: *tái cơ cấu, tái đầu tư, tái định cư, tái sản xuất*

tiểu + N = N: *tiểu bang, tiểu công nghệ, tiểu gia súc, tiểu khí hậu, tiểu loại, tiểu vương quốc*
 trưởng + N = N: *trưởng ban, trưởng phòng, trưởng thôn, trưởng tộc*
 tối + A = A: *tối đại đa số, tối thông minh (cần khảo sát tiếp)*
 vô + N = A: *vô chủ, vô đạo, vô đạo đức, vô gia cư, vô nhân đạo, vô thân, vô văn hoá*
 vô + V = A: *vô can, vô địch, vô học,...*
 vô + V = P: *vô kể, vô luận*

5.2. Danh sách các yếu tố đứng sau

N + hoá = V: *lao động hoá, công nông hoá, trí thức hoá*
 N + kiều = N: *Ấn kiều, Hoa kiều, Việt kiều*
 N + trưởng = N: *đại đoàn trưởng, phân viện trưởng, tiểu đoàn trưởng,*
 V + viên = N: *cộng sự viên, lập trình viên, điều tra viên*
 N + viên = N: *công an viên*

6. Tổ hợp có tính thành ngữ, quán ngữ

6.1. Danh sách các đơn vị thành ngữ

anh hùng áo vải	hoá chính vi linh	quanh đi quẩn lại
ăn càn nói rõ	huynh đệ chi bang	quân nào tướng nấy
ăn cơm chúa múa tối ngày	hương lạnh khói tàn	sĩ nông công thương
ăn đói mặc rách	hữu tiến vô thoái	suy đi nghĩ lại
buốt như kim châm	khoanh tay chờ chết	tan nhà nát cửa
bụng chứa vượt mặt	lai vô ảnh khứ vô tung	tán gia bại sản
bữa rau bữa cháo	lầu son gác tía	thâm sơn cùng cốc
chân mây ngọn sóng	mắt to hơn bụng	thiên kinh vạn quyển
chia ba xẻ bảy	một cổ đôi lòng	thuật kỳ phép lạ
chủ quan khinh địch	một mất một còn	tiền nghìn bạc vạn
có thực mới vực được đạo	một sống một chết	tối mù tối mịt
con Lạc cháu Hồng	muốn gì được vậy	trai tứ chiếng gái giang hồ
cứu khổ cứu nạn	muru sâu kẻ giỏi	trời cao đất dày
dân chi phụ mẫu	năm chùng mười hoạ	trời xanh nước biếc
dầu sương dãi nắng	người quen kẻ thuộc	trường xuân bất lão
đánh ngay thắng ngay	như muối bỏ bể	tuổi già sức yếu
đi nắng về mưa	nhức đầu sổ mũi	tư thù tư oán
đồng chu cộng tế	nhức như búa bổ	vay quanh mượn quản
đủ ăn đủ mặc	no đói có nhau	vắt cam vứt xác
đường đi nước bước	nổi như cồn	vợ đẹp con khôn
giết người cướp của	nước mất nhà tan	...

6.2. Danh sách các đơn vị quán ngữ

lễ với nghĩa
vợ với con
đáng chú ý là
mặt khác thì
nói cho cùng
nói một tiếng
nói tóm lại = tóm lại
v.v...

7. Tên riêng

* Tên người, tên địa danh, tên tổ chức được coi là một đơn vị từ vựng: tách theo quy định tách từ thông thường, riêng danh từ riêng thì gộp làm một.

- Tên tổ chức:

báo - Tuổi trẻ
Công ty - Cao su - Đồng Nai
Điện lực - Bến Tre

- Tên địa danh:

xã - Xuân Thanh
huyện - Long Khánh
tỉnh - Đồng Nai
Nông trường - Cẩm Đường
TP. – HCM
sông - Nhơn Mỹ
chợ - Phương Lâm

- Tên người:

bạn đọc - Nguyễn Hữu Ngọc Anh
bạn đọc - Nguyễn Thừa Nghiệp
Bạn đọc - Phan Văn Chiến

8. Ngày – tháng – năm, số – chữ số – kí hiệu

8.1. Ngày – tháng – năm.

- Giữ nguyên cả khối với các dạng (trong ngoặc không tính đến):

30-4-1975; 30-04-1975; 30-4-75; 30-04-75

(Ngày) 1-6; 01-06;

(Quốc khánh) 2-9; 02-09

- Tách thành từng đơn vị số, dấu, chữ như quy định thông thường:

tháng / 6 / – / 2003, Năm / 1997

8.2. Số – chữ số – kí hiệu

Các chuỗi này thường xử lý được dưới dạng biểu thức chính quy.

- Công thức hoá học, biểu thức toán học giữ nguyên cả khối:

$H + O_2 = H_2O$; $100 - x + 5 = 50$; $x - 23 < 23$

- Biểu hiện liên tục một con số chính xác **bằng số** (có dấu chấm: 1.500, không có dấu chấm 23000, VII, hay có dấu cách 1 000) hoặc **bằng chữ** (VD: hai mươi vạn, hai mươi phẩy hai, ba phần tư).

- Biểu hiện đặc biệt cả số và kí hiệu một cách liên tục (không có dấu cách) như: 19g25, 50%, 20ha.

- Biểu hiện **hỗn hợp** cả số và chữ thì tách riêng từng phần:

60	hai mươi nghìn	2	121, 8
phần trăm	tấn	-	tỉ
	rưỡi	3	
		triệu	
100			
phần			
100			

9. Dấu câu

- Tách riêng toàn bộ các loại dấu câu.

10. Từ tiếng nước ngoài

- Với các từ, thuật ngữ, khái niệm thì tách theo từng khối kí tự viết liền.

- Đối với tên riêng (tên người, tên địa danh) viết theo dạng đầy đủ thì tách theo mục 7.

- Trường hợp tên người và tên đệm viết tắt thì vẫn giữ nguyên cả khối:

V. E. Lênin,

11. Chữ viết tắt

- Tách theo từng khối kí tự viết liền:

ADSL, CNXH

- Chữ viết tắt là một bộ phận của tên riêng thì xử lý giống như tên riêng:

Đại học KHXH&NV Hà Nội, Cty TNHH Rạng Đông

Một số lưu ý khi thực hiện công việc tách từ vòng 2:

1) Các cụm từ chỉ ngày, tháng, năm một cách chính xác vòng 1 đã gộp thì bây giờ tách ra theo quy định trên.

2) Các trường hợp có dấu cách trước và sau dấu chấm ở chữ số (13 . 000), trước và sau dấu phẩy ở số thập phân: (3 , 5) thì xoá dấu cách và để thành 1 đơn vị.

3) Có những biểu thức có dấu ‘/’ như phân số, nhưng thực chất không phải thì phải tách: VD ở file 1019.txt: chỉ – có – 1 – / – 7 – khu – đã – khởi công – xây dựng (1/7 ở đây đọc là “một trên bảy” hoặc “một trong bảy” chứ không đọc “một phần bảy”); có những biểu thức có dấu ‘,’ như số thập phân, nhưng thực chất không phải thì phải tách: VD ở file 1019.txt: phát triển – 1 . 447 – km – đường ống – cấp – 1 – , – 2 – , – 3 (dấu phẩy ở đây là dấu câu chứ không phải dấu trong số thập phân).

3) Các trường hợp có dấu cách sau các chữ viết tắt (TP .) thì xoá dấu cách và để thành một đơn vị.

4) Hiện tượng nhập nhằng về nghĩa: Rất nhiều trường hợp từ được tách đúng về mặt hình thức, nhưng sai về nghĩa trong ngữ cảnh cụ thể, đòi hỏi người tách từ phải nhận ra và sửa lại cho đúng:

- rút xuống sông vì / cầu sập	- cử phóng viên làm / tin	- về vụ cháy / chợ Phương Lâm	- anh giật / dây cầu cứu	- Hầm đi / sâu vào lòng núi	- nhận được / cái lắc đầu
-------------------------------	---------------------------	-------------------------------	--------------------------	-----------------------------	---------------------------

- thừa ủy nhiệm kiến trúc sư trưởng TP	- Trường hợp căn / số 365 PNL được cấp GPXD	- do Công ty TNHH Hai Thành làm / chủ đầu tư		
--	---	--	--	--

5) Các lưu ý khác:

- Khi gặp những trường hợp rất khó xác định hoặc khi quyết định những đơn vị từ không có trong Từ điển thì phải ghi chú lại, thảo luận để tìm cách giải quyết thống nhất trong nhóm và đảm bảo tính nhất quán trong tư liệu.

- Có những trường hợp vòng 1 gộp nhưng bây giờ thấy nên tách ra thì đúng hơn: **một / nửa; mà / còn.**

- Có những trường hợp vòng 1 tách nhưng bây giờ thấy nên tách ra thì đúng hơn: **nhà ở (phân biệt với nhà xưởng).**