CÁC THAM SỐ TH À NH L Ậ P KHO NGỮ LI Ệ U MÁY TÍNH

TS. Lâm Thị Ho à Bình

\*1

Tóm tắt: Suốt gần một thế kỷ qua, các kho ngữ liệu (KNL) đã và đang được xây dựng và sử dụng rộng rãi

trong nghiên cứu ngôn ngữ cũng như giảng dạy ngoại ngữ. Cùng với sự phát triển của khoa học công nghệ,

các ngành khoa học tự nhiên và xã hội ngày càng xích lại gần nhau, tương tác và bổ sung cho nhau. Các kho

ngữ liệu hiện đại ngày nay được thành lập là kết quả của công nghệ với những cải tiến trong thu thập, chú

giải và phân tích dữ liệu ngôn ngữ. Bài viết trình bày về các tham số thành lập nên kho ngữ liệu hiện đại và

cho thấy sự kết hợp giữa các tham số này trong việc thành lập kho ngữ liệu.

Từ khóa: Ngôn ngữ học ngữ liệu; tham số; tập hợp mục tiêu; chú giải; phân tích.

Abstract: For nearly a century, linguistic corpora have been built and utilized widely in linguistic researches

as well as teaching foreign languages. Along with technology development, the gap between natural science

and social science has become narrower with more and more interactions and collaborations among various

disciplines. The involvement of technology results in modern corpora with the improvement in collecting,

annotating and analyzing linguistic data. The report demonstrates three parameters in building a modern

linguistic corpus as well as their combination in a corpus formation.

Key words: corpus; parameters; target population; annotation; analysis.