Steps toward global interoperability for language resources

D. Terence Langendoen

Professor Emeritus of Linguistics, University of Arizona & Program Officer for Linguistics, National Science Foundation

Scope of paper

- Language resources
  - Language descriptions including glossed texts, treebanks, lexicons, grammars, etc.
- Interoperability
  - For feature-based analyses and descriptions (FADs), i.e. those making use of features (attribute-value pairs) and structured objects comprised of features.

Outline

- Features
  - Feature systems
  - Sharable feature systems
- Feature structures
  - Feature-structure systems
  - Sharable feature-structure systems

A feature system (F-system) F_A consists of:

1. a set V_A of two or more features for a particular attribute A that are distinguished by their values, and
2. the subsumption relation $\sqsubseteq$, a partial ordering over the members of V_A.
   - My discussion here is limited to attributes with "symbolic" values (Langendoen & Simons 1995), at least two of which are atoms as described in the next slide.
   - F_A is an implication structure in the sense of Koslow (1992).

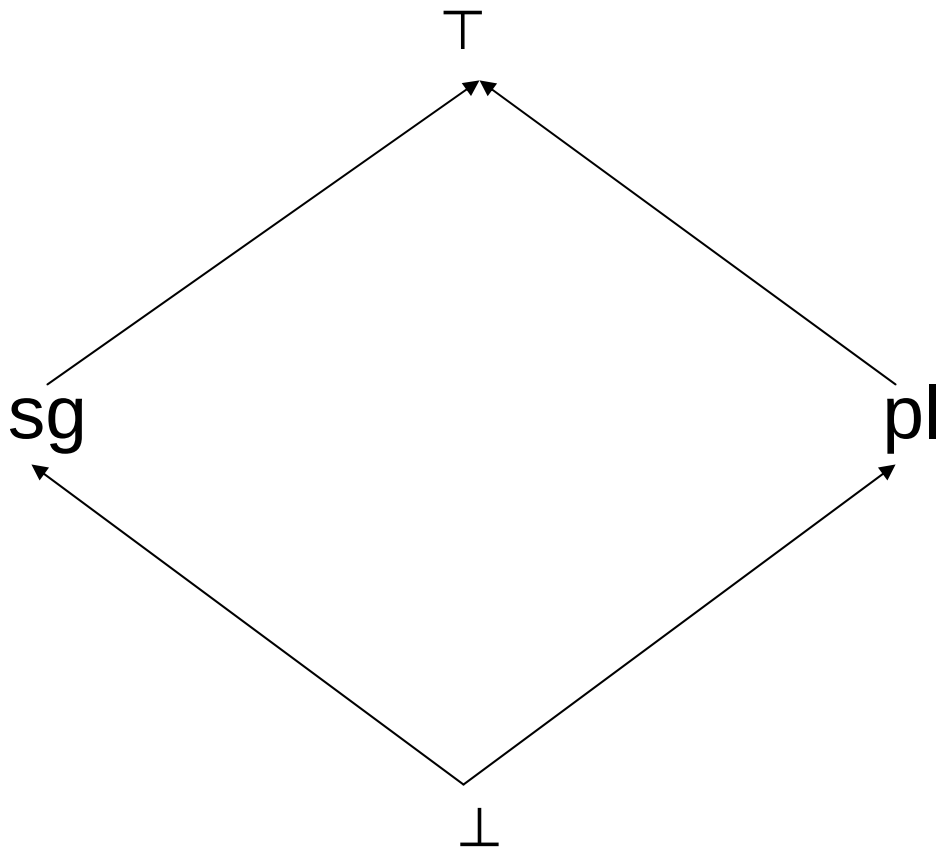
Binary F-systems

- The simplest F-systems have exactly two features that do not subsume each other, and are not the disjunctions of any other features in their systems. Such features are atoms, e.g.
  1. sg = [Number Singular]
  2. pl = [Number Plural]
- Such a “binary” F-system may also contain the disjunction and the conjunction of the atoms, e.g.
  3. sg|pl = [Number Singular|Plural] = ⊤ “top” or “verum”
  4. sg&pl = [Number Singular&Plural] = ⊥ “bottom” or “falsum”

Example binary F-systems

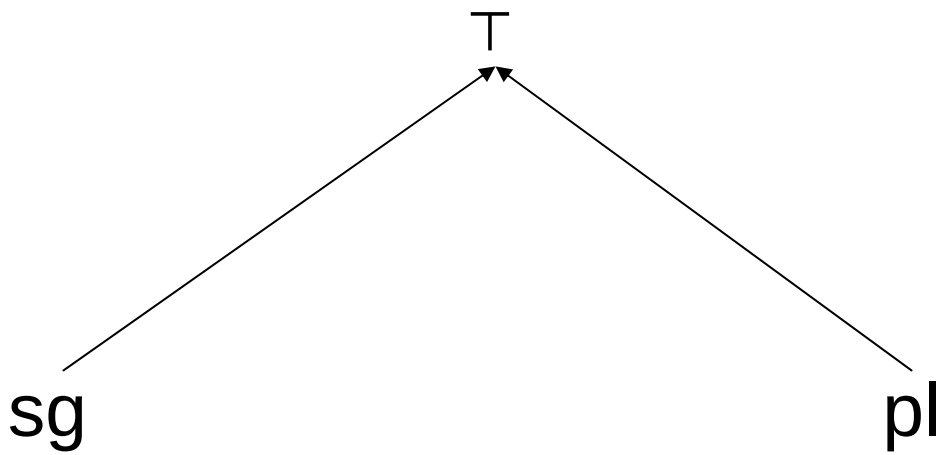
- The next three slides diagram the F-systems F_{N4}, F_{N3} and F_{N2}, for the Number feature sets:
  - V_{N4} = {sg, pl, $\top$, $\bot$}
  - V_{N3} = {sg, pl, $\top$} (omitting $\bot$)
  - V_{N2} = {sg, pl} (omitting $\top$, $\bot$)
  - –In these diagrams, the arcs derivable from the reflexivity and transitivity of the subsumption relation are not shown.
  - –F_{N4} represents the closure of a two-atom Number system with respect to conjunction, disjunction and negation, and so is a **maximal** Number system with two atoms.

F_{N4}



Read the subsumption arcs downwards for conjunction, and upwards for disjunction. The negation of a feature is the feature appearing in its "reflection" using the x or y axis as a mirror, so that *sg* and *pl* are negations of each other (i.e. *sg* = ~*pl* and *pl* = ~*sg*), as are ⊤ and ⊥.

F_{N3}



F_{N3} is a nonclassical F-system in which ⊤ has no negation and there is no conjunction of *sg* with *pl*.

F_{N2}

sg pl

F_{N2} is a nonclassical F-system in which there is no conjunction or disjunction of *sg* with *pl*. However, *sg* and *pl* continue to be each other's negations.

Interoperability for binary F-systems

- Achieving interoperability for FADs using binary F-systems requires agreement concerning:
  - attribute names,
  - identity of the two atomic values for such features, and
  - interpretation of ⊤ if used.
    - In such F-systems, ⊤ may represent either:
      - underspecification of a binary feature's value, or
      - uncertainty about that feature's value.

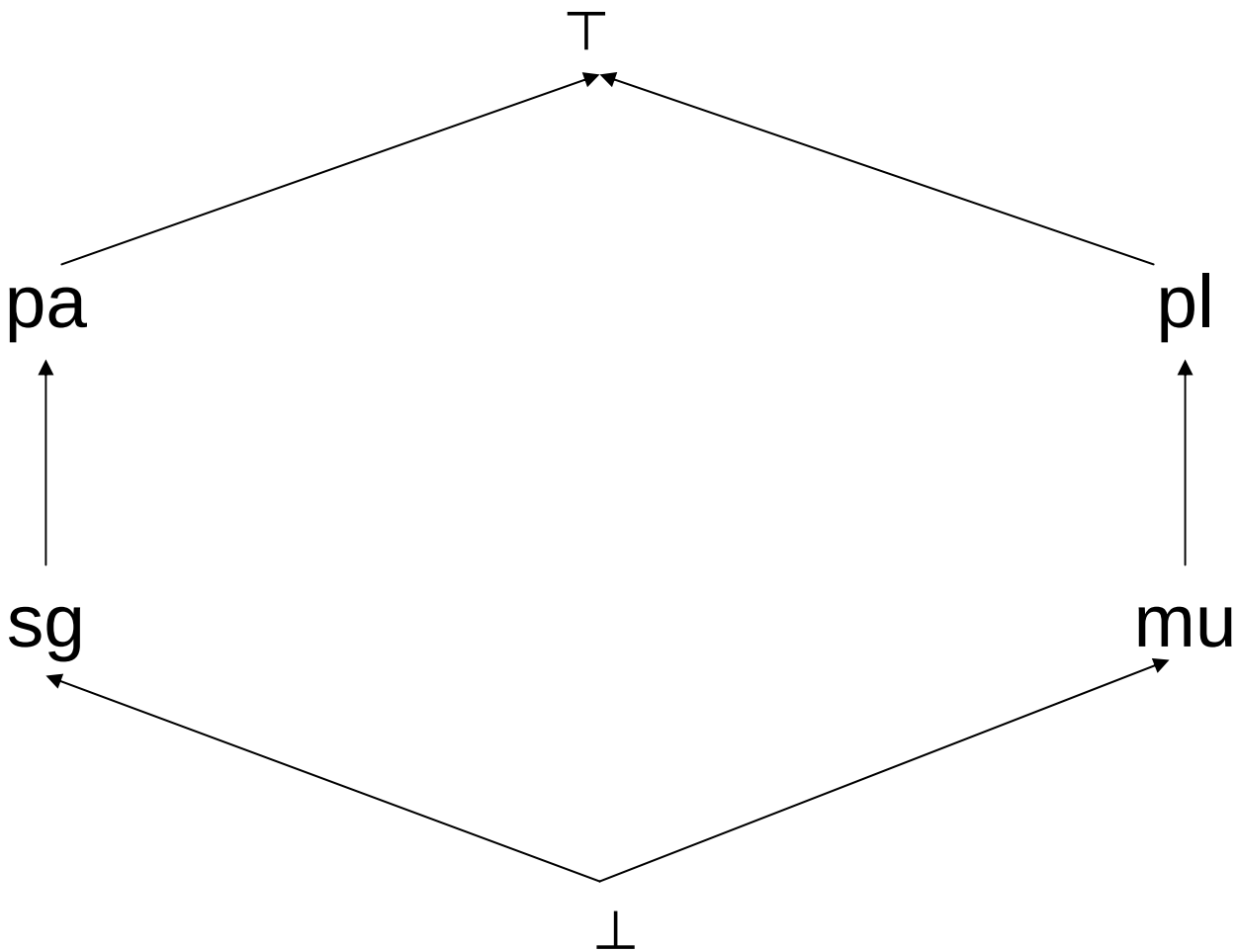
Outline

- Features
  - Feature systems
  - Sharable feature systems
- Feature structures
  - Feature-structure systems
  - Sharable feature-structure systems

Binary F-subsystems of larger F-systems

- Suppose the binary F-system F_{N4a} is proposed for some language, which is just like F_{N4}, but with *pa* = [Number Paucal] and *mu* = [Number Multal] replacing *sg* and *pl* respectively, and it is agreed that *sg*, *pl*, *pa*, *mu*, ⊤ and ⊥ are all possible values for the Number feature across FADs.
  - Then there is a larger F-system of which F_{N4} and F_{N4a} are binary F-subsystems.
  - Merging these subsystems, assuming that *sg* ⊑ *pa* and *mu* ⊑ *pl*, yields the F-system F_{N6} diagrammed in the next slide.

F_{N6}

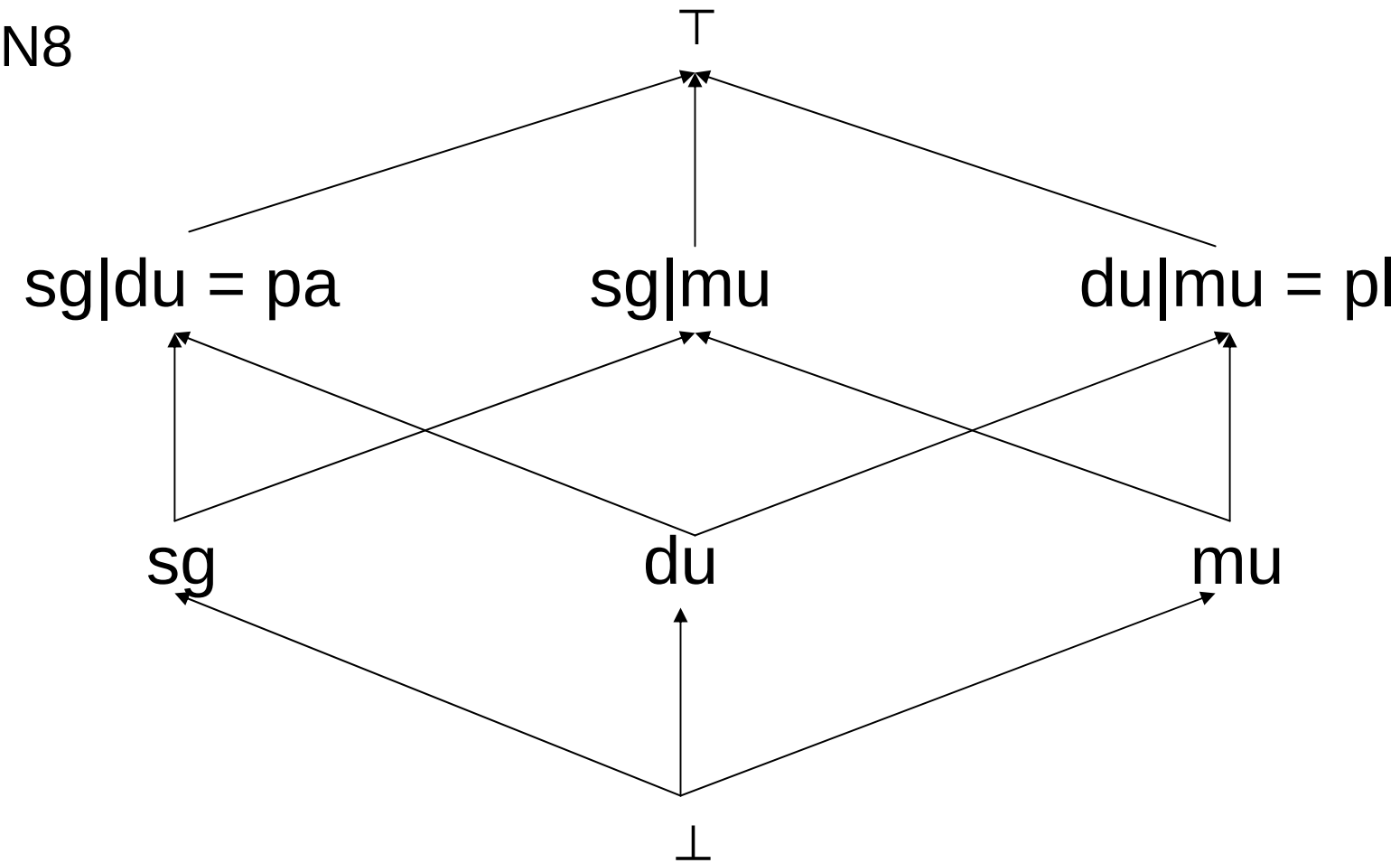


F_{N6} is not a binary F-system because it contains more than one pair of atoms, in fact four: {sg, pl}, {pa, mu}, {pa, pl} and {sg, mu}.

Continuing the process

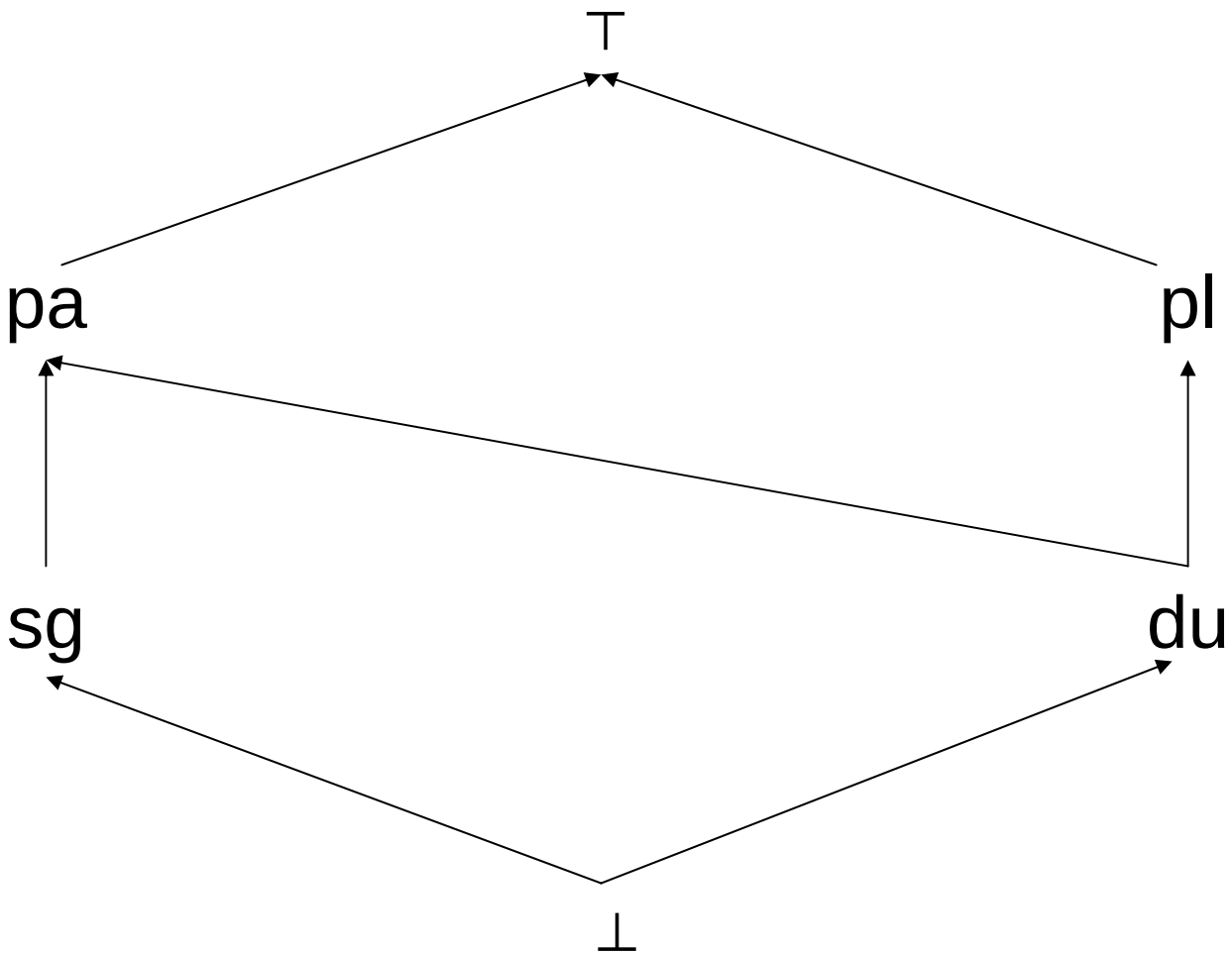
- Next, suppose an F-system is proposed and accepted, which is a subsystem of the maximal ternary F-system F_{N8} over the atoms *sg*, *du* = [Number Dual], and *mu*, as diagrammed in the next slide.
- Then if the identification of *pa* with *sg*|*du*, and *pl* with *du*|*mu* is also accepted, F_{N6} together with all the other Number F-systems considered so far, and many others, are all F-subsystems of F_{N8}.

F_{N8}



F_{N6a}, which differs from F_{N6} only in the occurrence of *du*, where F_{N6a} has *mu*, is diagrammed in the next slide.

F_{N6a}



F_{N6a} is a nonclassical F-system in which the laws of double negation and excluded middle fail, since ~~*du* = *pl*, not *du*, and *du*|~*du* = *pa*, not ⊤.

Sharable F-systems (SF-systems)

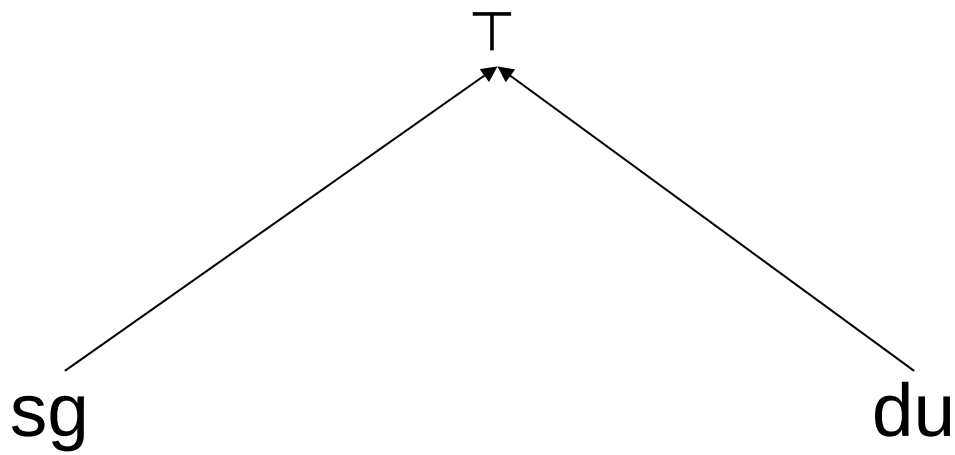
- An F-system F_{Amax}, such as F_{N8}, obtained by:
  - merging the F-systems of a language attribute A, and
  - forming the closure of over the atoms of the result

  does not necessarily represent a system associated with any FAD.
- Rather it is to be thought of as a “sharable” F-system (SF-system) that has the capacity to contain every FAD’s F-system over A as a subsystem.

Size and structure of SF-systems

- An SF-system with *n* atoms contains 2^n features, and converges on 2^{2^n} F-subsystems in the limit.
  - F_{N4} has 2 atoms, 4 features and 4 F-subsystems.
  - F_{N8} has 3 atoms, 8 features and 210 F-subsystems.
  - As *n* increases, perhaps some features with disjunctive values can be deprecated, sharply reducing the number of “linguistically significant” F-subsystems.
  - “Defective” F-subsystems, as in the next slide, may also be deprecated.

F_{N3a}



F_{N3a} is a defective F-subsystem of F_8 because the disjunction of its atoms $\neq \top$ **in F_8**. On the other hand F_{N3}, diagrammed in slide 8, is not defective because the disjunction of its atoms $= \top$ in F_8.

Choice of F-subsystems

- Analysts' preference for relatively unstructured ("flat") F-subsystems, as expressed, for example at the 2005 E-MELD workshop, does not imply that SF-systems should also be relatively unstructured.
- F-subsystems of any desired degree of flatness are easily obtained from a highly structured SF-system; an example is given in the next slide.

F_{N3b}

sg du mu

F_{N3b} is a perfectly “flat” F-subsystem of the SF-system F_{N8}.

SF-systems provide interoperability over FS-subsystems

- They provide an explicit basis of comparison of F-systems.
  - Are they the same?
  - If not, exactly how do they differ?
- They provide support for structured queries and inferencing.
- They enable one to predict consequences of changes to analyses.

Outline

- Features
  - Feature systems
  - Sharable feature systems
- Feature structures
  - Feature-structure systems
  - Sharable feature-structure systems

A feature-structure system (FS-system) $FS_{A\times B\ \ldots}$ consists of:

1. A subset $W_{A\times B\ldots}$ of the union of $\perp$ with the Cartesian product $V_{A\times B} = V_A\times V_B\times\ldots$ of the features of two or more F-systems F_A, F_B, … that all lack $\perp$;
2. the subsumption relation $\sqsubseteq$, a partial ordering over the members of $W_{A\times B\ldots}$.
   - $a_1\times b_1 \sqsubseteq a_2\times b_2$ in $W_{A\times B}$ iff $a_1 \sqsubseteq a_2$ in F_A and $b_1 \sqsubseteq b_2$ in F_B.
   - $\perp \sqsubseteq a\times b$ for every a in V_A and b in V_B.

Doubly binary FS-systems

- The simplest FS-systems result from taking the Cartesian product of the members of two binary F-systems.
- If the two F-systems are maximal (except for $\perp$), $\perp$ is included, and $W_{A\times B} = V_{A\times B}$, the resulting maximal double binary FS-system contains $2\times2 = 4$ atoms and $2^4 = 16$ FSs.
- The smallest double binary FS-system has 2 atoms and 2 FSs.

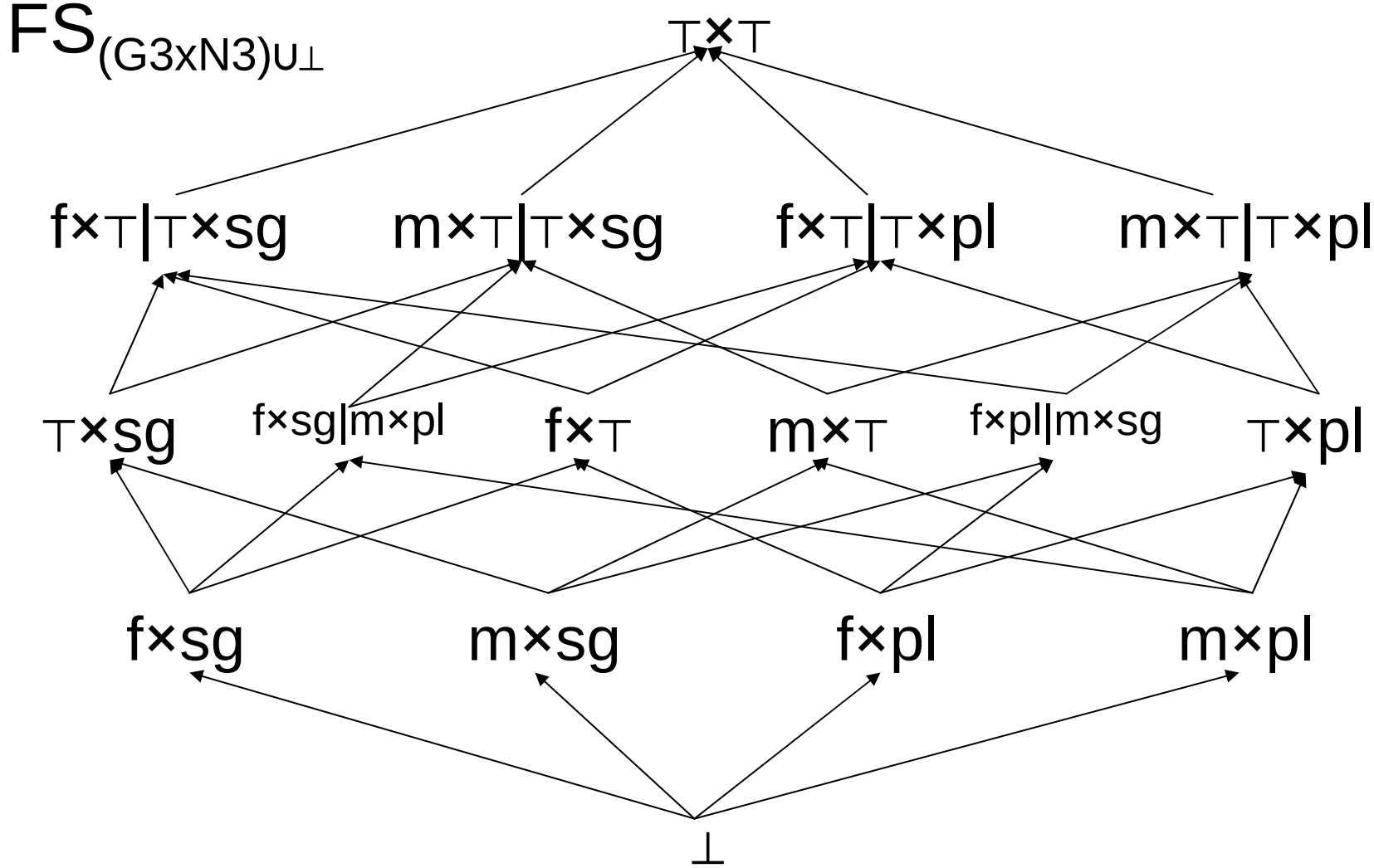
Examples of doubly binary FS-systems

- The next three slides diagram the doubly binary FS-systems $FS_{(G3xN3)U\perp}$, $FS_{(G3xN3)a}$, $FS_{(G3xN3)b}$ where:

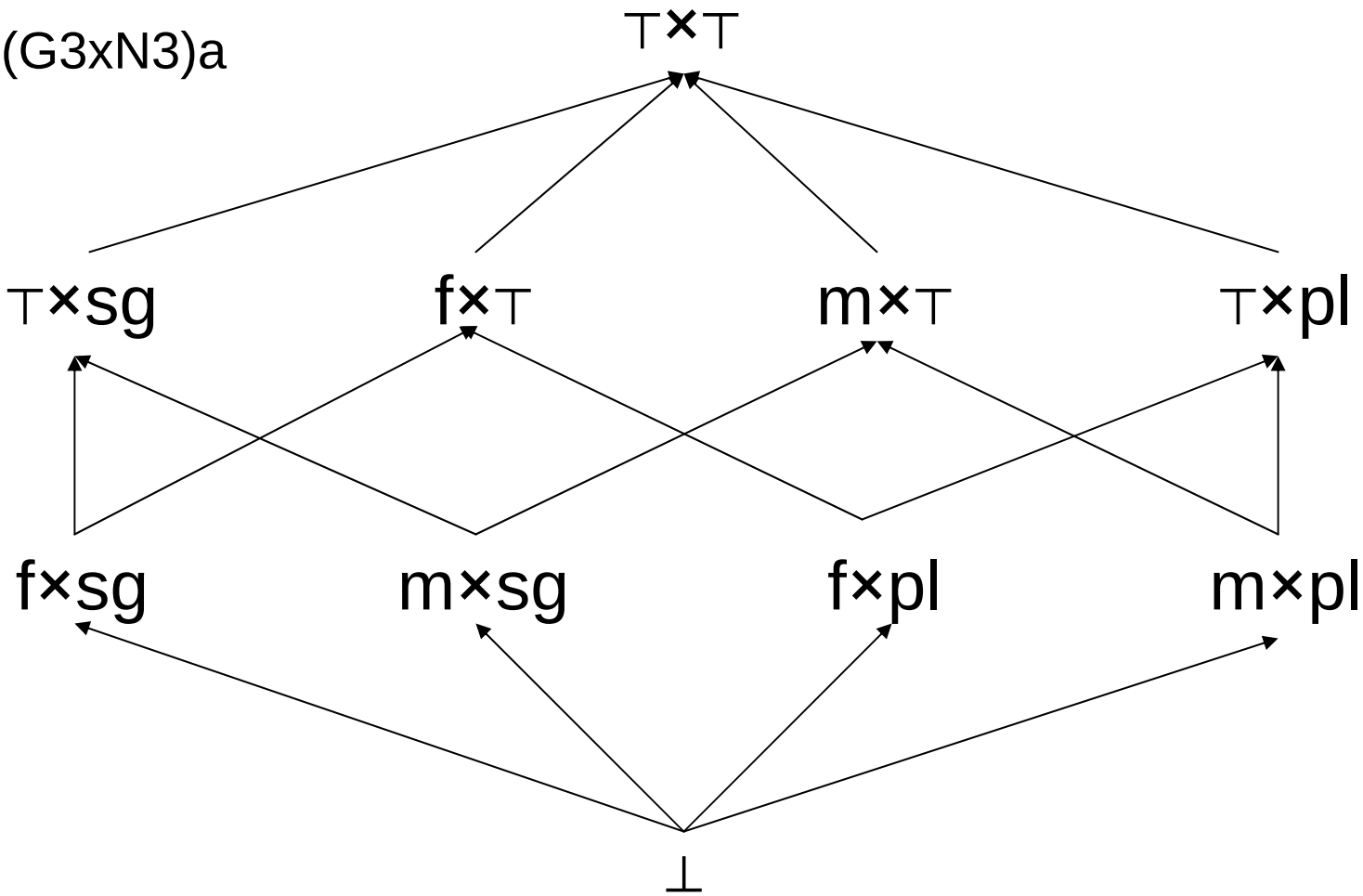
  V_{G3} = {f = [Gender Feminine], m = [Gender Masculine], $\top$}

  V_{N3} = {sg, pl, $\top$}

  - $FS_{(G3xN3)U\perp}$ is a maximal doubly binary FS-system.
  - $FS_{(G3xN3)a}$ is a subsystem of $FS_{(G3xN3)U\perp}$ in which all the explicitly disjunctive FSs have been removed.
  - $FS_{(G3xN3)b}$ is a subsystem of $FS_{(G3xN3)U\perp}$ containing two atoms only.


FS(G3xN3)∪⊥
⊤×⊤
f×⊤|⊤×sg
m×⊤|⊤×sg
f×⊤|⊤×pl
m×⊤|⊤×pl
⊤×sg
f×sg|m×pl
f×⊤
m×⊤
f×pl|m×sg
⊤×pl
f×sg
m×sg
f×pl
m×pl
⊥

$FS_{(G3xN3)a}$



$FS_{(G3xN3)b}$

f×sg

m×⊤|⊤×pl = ~(f×sg)

Paradigmatic structures as FS-systems

- A linguistic paradigm can be described as an FS-system.
  - For example, an inflectional paradigm for:

    3 atomic genders,

    2 atomic numbers and

    4 atomic cases

    is a subsystem of a maximal FS-system containing 2^{24} FSs.
  - The number of possible such subsystems approximates $2^{2^{24}}$!

Outline

- Features
  - Feature systems
  - Sharable feature systems
- Feature structures
  - Feature-structure systems
  - Sharable feature-structure systems

Sharable FS-systems (SFS-systems)

- An SFS-system can be obtained by taking the Cartesian product of the features of its component SF-systems, other than $\perp$, and re-introducing $\perp$ at the end. $FS_{(G3xN3)U\perp}$ would be an SFS-system if its components were SF-systems.
- As we've already seen, such SFS-systems can be very large, and the number of FS-subsystems is exponentially larger.
  - To make the use of SFS-systems practical, a means for massively pruning them would have to be developed and accepted.

SFS-systems provide interoperability over FS-subsystems

- The virtues enumerated in slide 22 for SF-systems carry over to SFS-systems.
- To the extent that linguistic analyses are expressible with FS-subsystems of the sort described here, SFS-systems can be used to compare, query, and modify those analyses.

References, acknowledgment and disclaimer

- References

Koslow, Arnold (1992) *A Structuralist Theory of Logic*. Cambridge University Press.

Langendoen, D. Terence & Gary F. Simons (1995) A rationale for the TEI recommendations for feature-structure markup. *Computers and the Humanities* 29: 191-209.

- Acknowledgment

I read an earlier version of this paper entitled "Elementary reasoning with features“ on 11 Dec 2007 at the University of Surrey. I thank the members of that audience for many helpful suggestions.

- Disclaimer

This material is based in part upon work supported while I was serving at the National Science Foundation. Any opinion and conclusions are mine and do not necessarily reflect the views of the National Science Foundation.